Gradient-Based Adaptive Stochastic Search for Non-Differentiable Optimization

Enlu Zhou

Department of Industrial & Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, IL 61801, enluzhou@illinois.edu

Jiaqiao Hu

Department of Applied Mathematics and Statistics, Stony Brook University, NY 11794, jqhu@ams.sunysb.edu

This version: October 22, 2012

ABSTRACT

In this paper, we propose a stochastic search algorithm for solving general optimization problems with little structure. The algorithm iteratively finds high quality solutions by randomly sampling candidate solutions from a parameterized distribution model over the solution space. The basic idea is to convert the original (possibly non-differentiable) problem into a differentiable optimization problem on the parameter space of the parameterized sampling distribution, and then use a direct gradient search method to find improved sampling distributions. Thus, the algorithm combines the robustness feature of stochastic search from considering a population of candidate solutions with the relative fast convergence speed of classical gradient methods by exploiting local differentiable structures. We analyze the convergence and converge rate properties of the proposed algorithm, and carry out numerical study to illustrate its performance.

1. Introduction

We consider global optimization problems over real vector-valued domains. These optimization problems arise in many areas of importance and can be extremely difficult to solve due to the presence of multiple local optimal solutions and the lack of structural properties such as differentiability and convexity. In such a general setting, there is little problem-specific knowledge that can be exploited in searching for improved solutions, and

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
22 OCT 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Gradient-Based Adaptive Stochastic Search for Non-Differentiable Optimization	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Illinois at Urbana-Champaign,Department of Industrial & Enterprise Systems Engineering,Urbana,IL,61801**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
submitted

14. ABSTRACT

In this paper, we propose a stochastic search algorithm for solving general optimization problems with little structure. The algorithm iteratively nds high quality solutions by randomly sampling candidate solutions from a parameterized distribution model over the solution space. The basic idea is to convert the original (possibly non-di erentiable) problem into a di erentiable optimization problem on the parameter space of the parameterized sampling distribution, and then use a direct gradient search method to nd improved sampling distributions. Thus, the algorithm combines the robustness feature of stochastic search from considering a population of candidate solutions with the relative fast convergence speed of classical gradient methods by exploiting local di erentiable structures. We analyze the convergence and converge rate properties of the proposed algorithm, and carry out numerical study to illustrate its performance.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**34**	

it is often the case that the objective function can only be assessed through the form of "black-box" evaluation, which returns the function value for a specified candidate solution.

An effective and promising approach for tackling such general optimization problems is stochastic search. This refers to a collection of methods that use some sort of randomized mechanism to generate a sequence of iterates, e.g., candidate solutions, and then use the sequence of iterates to successively approximate the optimal solution. Over the past years, various stochastic search algorithms have been proposed in literature. These include approaches such as simulated annealing [10], genetic algorithms [7], tabu search [6], pure adaptive search [28], and sequential Monte Carlo simulated annealing [29], which produce a sequence of candidate solutions that are gradually improving in performance; the nested partitions method [25], which uses a sequence of partitions of the feasible region as intermediate constructions to find high quality solutions; and the more recent class of model-based algorithms (see a survey by [30]), which construct a sequence of distribution models to characterize promising regions of the solution space.

This paper focuses on model-based algorithms. These algorithms typically assume a sampling distribution (i.e., a probabilistic model), often within a parameterized family of distributions, over the solution space, and iteratively carry out the two interrelated steps: (1) draw candidate solutions from the sampling distribution; (2) use the evaluations of these candidate solutions to update the sampling distribution. The hope is that at every iteration the sampling distribution is biased towards the more promising regions of the solution space, and will eventually concentrate on one or more of the optimal solutions. Examples of model-based algorithms include ant colony optimization [4, 3], annealing adaptive search (AAS) [22], probability collectives (PCs) [27], the estimation of distribution algorithms (EDAs) [14, 19], the cross-entropy (CE) method [23], model reference adaptive search (MRAS) [8], and the interacting-particle algorithm [17, 18]. The various model-based algorithms mainly differ in their ways of updating the sampling distribution. Recently, [9] showed that the updating schemes in some model-based algorithms can be viewed under a unified framework. The basic idea is to convert the original optimization problem into a sequence of stochastic optimization problems with differentiable structures, so that the distribution updating schemes in these algorithms can be equivalently transformed into the form of stochastic approximation procedures for solving the sequence of stochastic optimization problems.

Because model-based algorithms work with a population of candidate solutions at each iteration, they demonstrate more robustness in exploring the solution space as compared with their classical counterparts that work with a single candidate solution each time (e.g., simulated annealing). The main motivation of this paper is to integrate this robustness feature of model-based algorithms into familiar gradient-based tools from classical differentiable optimization to facilitate the search for good sampling distributions. The underlying idea is to reformulate the original (possibly non-differentiable) optimization problem into a *differentiable* optimization problem over the parameter space of the sampling distribution, and then use a direct gradient search method on the parameter space to solve the new

formulation. This leads to a natural algorithmic framework that combines the advantages of both methods: the fast convergence of gradient-based methods and the global exploration of stochastic search. Specifically, each iteration of our proposed method consists of the following two steps: (1) generate candidate solutions from the current sampling distribution; (2) update the parameters of the sampling distribution using a direct gradient search method. Although there are a variety of gradient-based algorithms that are applicable in step (2) above, in this paper we focus on a particular algorithm that uses a quasi-Newton like procedure to update the sampling distribution parameters. Note that since the algorithm uses only the information contained in the sampled solutions, it differs from the quasi-Newton method in deterministic optimization, in that there is an extra Monte Carlo sampling noise involved at each parameter updating step. We show that this stochastic version of quasi-Newton iteration can be expressed in the form of a generalized Robbins-Monro algorithm, and this in turn allows us to use the existing tools from stochastic approximation theory to analyze the asymptotic convergence and convergence rate of the proposed algorithm.

The rest of the paper is organized as follows. We introduce the problem setting formally in Section 2. Section 3 provides a description of the proposed algorithm along with the detailed derivation steps. In Section 4, we analyze the asymptotic properties of the algorithm, including both convergence and convergence rate. Some preliminary numerical study are carried out in Section 5 to illustrate the performance of the algorithm. Finally, we conclude this paper in Section 6. All the proofs are contained in the Appendix.

2. Problem Formulation

Consider the maximization problem

$$x^* \in \arg\max_{x \in \mathcal{X}} H(x), \quad \mathcal{X} \subseteq \mathbb{R}^n. \tag{1}$$

where the solution space $\mathcal{X}$ is a nonempty compact set in $\mathbb{R}^n$, and $H : \mathcal{X} \to \mathbb{R}$ is a real-valued function. Denote the optimal function value as H^*, i.e., there exists an x^* such that $H(x) \leq H^* \triangleq H(x^*)$, $\forall x \in \mathcal{X}$. Assume that H is bounded on $\mathcal{X}$, i.e., $\exists H_{lb} > -\infty,\ H_{ub} < \infty$ s.t. $H_{lb} < H(x) < H_{ub}$, $\forall x \in \mathcal{X}$. We consider problems where the objective function $H(x)$ lacks nice structural properties such as differentiability and convexity and could have multiple local optima.

Motivated by the idea of using a sampling distribution/probabilistic model in model-based optimization, we let $\{f(x;\theta)|\theta \in \Theta \subseteq \mathbb{R}^d\}$ be a parameterized family of probability density functions (pdfs) on $\mathcal{X}$ with Θ being a parameter space. Intuitively, this collection of pdfs can be viewed abstractly as probability models characterizing our knowledge or belief of the promising regions of the solution space. It is easy to see that

$$\int H(x) f(x;\theta) dx \leqslant H^*, \quad \forall \theta \in \Theta.$$

In this paper, we simply write $\int$ with the understanding that the integrals are taken over $\mathcal{X}$. Note that the equality on the righthand side above is achieved whenever there exists an optimal parameter under which the parameterized probability distribution will assign all of its probability mass to a subset of the set of global optima of (1). Hence, one natural idea to solving (1) is to transform the original problem into an expectation of the objective function under the parameterized distribution and try to find the best parameter θ^* within the parameter space Θ such that the expectation under $f(x, \theta^*)$ can be made as large as possible, i.e.,

$$\theta^* = \arg\max_{\theta \in \Theta} \int H(x) f(x; \theta) dx. \tag{2}$$

So instead of considering directly the original function $H(x)$ that is possibly non-differentiable and discontinuous in x, we now consider the new objective function $\int H(x) f(x; \theta) dx$ that is continuous on the parameter space and usually differentiable with respect to θ. For example, under mild conditions the differentiation can be brought into the integration to apply on the p.d.f. $f(x; \theta)$, which is differentiable given an appropriate choice of the distribution family such as an exponential family of distributions.

The formulation of (2) suggests a natural integration of stochastic search methods on the solution space $\mathcal{X}$ with gradient-based optimization techniques on the continuous parameter space. Conceptually, that is to iteratively carry out the following two steps:

1. Generate candidate solutions from $f(x; \theta)$ on the solution space $\mathcal{X}$.

2. Use a gradient-based method for the problem (2) to update the parameter θ.

The motivation is to speed up stochastic search with a guidance on the parameter space, and hence combine the advantages of both methods: the fast convergence of gradient-based methods and the global exploration of stochastic search methods. Even though problem (2) may be non-concave and multi-modal in θ, the sampling from the entire original space $\mathcal{X}$ compensates the local exploitation along the gradient on the parameter space. In fact, our algorithm developed later will automatically adjust the magnitude of the gradient step on the parameter space according to the global information, i.e., our belief about the promising regions of the solution space.

For algorithmic development later, we introduce a shape function $S_\theta : \mathbb{R} \to \mathbb{R}^+$, where the subscript θ signifies the possible dependence of the shape function on the parameter θ. The function S_θ satisfies the following conditions:

(a) For every θ, $S_\theta(y)$ is nondecreasing in y and bounded from above and below for bounded y, with the lower bound being away from zero. Moreover, for every fixed y, $S_\theta(y)$ is continuous in θ;

(b) The set of optimal solutions $\{\arg\max_{x \in \mathcal{X}} S_\theta(H(x))\}$ is a non-empty subset of $\{\arg\max_{x \in \mathcal{X}} H(x)\}$, the set of optimal solutions of the original problem (1).

Therefore, solving (1) is equivalent to solving the following problem

$$x^* \in \arg\max_{x\in\mathcal{X}} S_\theta(H(x)). \tag{3}$$

The main reason of introducing the shape function S_θ is to ensure positivity of the objective function $S_\theta(H(x))$ under consideration, since $S_\theta(H(x))$ will be used to form a probability density function later. Moreover, the choice of S_θ can also be used to adjust the trade-off between exploration and exploitation in stochastic search. One choice of such a shape function, similar to the level/indicator function used in the CE method and MRAS, is

$$S_\theta(H(x)) = (H(x) - H_{lb})\frac{1}{1+e^{-S_0(H(x)-\gamma_\theta)}}, \tag{4}$$

where S_0 is a large positive constant, and γ_θ is the $(1-\rho)$-quantile

$$\gamma_\theta \triangleq \sup_l \{l : P_\theta\{x \in \mathcal{X} : H(x) \geq l\} \geq \rho\},$$

where P_θ denotes the probability with respect to $f(\cdot;\theta)$. Notice that $1/(1+e^{-S_0(H(x)-\gamma_\theta)})$ is a continuous approximation of the indicator function $I\{H(x) \geq \gamma_\theta\}$, this shape function S_θ essentially prunes the level sets below γ_θ. By varying ρ, we can adjust the percentile of elite samples that are selected to update the next sampling distribution: the bigger ρ, the less elite samples selected and hence more emphasis is put on exploiting the neighborhood of the current best solutions. Sometimes the function S_θ could also be chosen to be independent of θ, i.e., $S_\theta = S : \mathbb{R} \to \mathbb{R}^+$, such as the function $S(y) = \exp(y)$.

For an arbitrary but fixed $\theta' \in \mathbb{R}^d$, define the function

$$L(\theta;\theta') \triangleq \int S_{\theta'}(H(x))f(x;\theta)dx.$$

According to the conditions on S_θ, it always holds that

$$0 < L(\theta;\theta') \leq S_{\theta'}(H^*) \quad \forall\, \theta,$$

and the equality is achieved if there exists an optimal parameter such that the probability mass of the parameterized distribution is concentrated only on a subset of the set of global optima. Following the same idea that leads to (2), solving (3) and thus (1) can be converted to the problem of trying to find the best parameter θ^* within the parameter space by solving the following maximization problem:

$$\theta^* = \arg\max_{\theta\in\Theta} L(\theta;\theta'). \tag{5}$$

Same as problem (2), $L(\theta;\theta')$ may be nonconcave and multi-modal in θ.

3. Gradient-Based Adaptive Stochastic Search

Following the formulation in the previous section, we propose a stochastic search algorithm that carries out the following two steps at each iteration: let θ_k be the parameter obtained at the k^{th} iteration,

1. Generate candidate solutions from $f(x;\theta_k)$.
2. Update the parameter to θ_{k+1} using a quasi Newton's iteration for $\max_\theta L(\theta;\theta_k)$.

Assuming it is easy to draw samples from $f(x;\theta)$, then the main obstacle is to find expressions of the gradient and Hessian of $L(\theta;\theta_k)$ that can be nicely estimated using the samples from $f(x;\theta)$. To overcome this obstacle, we choose $\{f(x;\theta)\}$ to be an exponential family of densities defined as below.

Definition 1. *A family $\{f(x;\theta):\theta\in\Theta\}$ is an exponential family of densities if it satisfies*

$$f(x;\theta)=\exp\{\theta^T T(x)-\phi(\theta)\},\quad \phi(\theta)=\ln\left\{\int \exp(\theta^T T(x))dx\right\}. \tag{6}$$

where $T(x)=[T_1(x),T_2(x),\ldots,T_d(x)]^T$ is the vector of sufficient statistics, $\theta=[\theta_1,\theta_2,\ldots,\theta_d]^T$ is the vector of natural parameters, and $\Theta=\{\theta\in\mathbb{R}^d:\ |\phi(\theta)|<\infty\}$ is the natural parameter space with a nonempty interior.

Define the density function

$$p(x;\theta)\triangleq\frac{S_\theta(H(x))f(x;\theta)}{\int S_\theta(H(x))f(x;\theta)dx}=\frac{S_\theta(H(x))f(x;\theta)}{L(\theta;\theta)}. \tag{7}$$

With $f(\cdot;\theta)$ from an exponential family, we propose the following updating scheme for θ in step 2 above:

$$\theta_{k+1}=\theta_k+\alpha_k(\mathrm{Var}_{\theta_k}[T(X)]+\epsilon I)^{-1}\left(E_{p_k}[T(X)]-E_{\theta_k}[T(X)]\right), \tag{8}$$

where $\epsilon>0$ is a small positive number, $\alpha_k>0$ is the step size, E_{p_k} denotes the expectation with respect to $p(\cdot;\theta_k)$, and E_{θ_k} and Var_{θ_k} denote the expectation and variance taken with respect to $f(\cdot;\theta_k)$, respectively. The role of ϵI is to ensure the positive definiteness of $(\mathrm{Var}_{\theta_k}[T(X)]+\epsilon I)$ such that it can be inverted. The term $(E_{p_k}[T(X)]-E_{\theta_k}[T(X)])$ is an ascent direction of $L(\theta;\theta_k)$, which will be shown in the next section.

To implement the updating scheme (8), the term $E_{p_k}[T(X)]$ is often not analytically available and needs to be estimated. Suppose $\{x_1,\ldots,x_{N_k}\}$ are independent and identically distributed (i.i.d.) samples drawn from $f(x;\theta_k)$. Since

$$E_{p_k}[T(X)]=E_{\theta_k}\left[T(X)\frac{p(X;\theta_k)}{f(X;\theta_k)}\right],$$

we compute the weights $\{w_k^i\}$ for the samples $\{x_k^i\}$ according to

$$w_k^i \propto \frac{p(x_k^i;\theta_k)}{f(x_k^i;\theta_k)} \propto S_{\theta_k}(H(x_k^i)), \quad i = 1, \ldots, N_k,$$
$$\sum_{i=1}^{N} w_k^i = 1.$$

Hence, $E_{p_k}[T(X)]$ can be approximated by

$$\widetilde{E}_{p_k}[T(X)] = \sum_{i=1}^{N_k} w_k^i T(x_k^i). \tag{9}$$

Some forms of the function $S_{\theta_k}(H(x))$ have to be approximated by samples as well. For example, if $S_{\theta_k}(H(x))$ takes the form (4), the quantile γ_{θ_k} needs to be estimated by the sample quantile. In this case, we denote the approximation by $\widehat{S}_{\theta_k}(H(x))$, and evaluate the normalized weights according to

$$\widehat{w}_i^k \propto \widehat{S}_{\theta_k}(H(x_k^i)), \quad i = 1, \ldots, N_k.$$

Then the term $E_{p_k}[T(X)]$ is approximated by

$$\widehat{E}_{p_k}[T(X)] = \sum_{i=1}^{N_k} \widehat{w}_k^i T(x_k^i). \tag{10}$$

In practice, the variance term $\mathrm{Var}_{\theta_k}[T(X)]$ in (8) may not be directly available or could be too complicated to compute analytically, so it also often needs to be estimated by samples:

$$\widehat{\mathrm{Var}}_{\theta_k}[T(X)] = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} T(x_k^i) T(x_k^i)^T - \frac{1}{N_k^2 - N_k} \left(\sum_{i=1}^{N_k} T(x_k^i) \right) \left(\sum_{i=1}^{N_k} T(x_k^i) \right)^T \tag{11}$$

The expectation term $E_{\theta_k}[T(X)]$ can be evaluated analytically in most cases. For example, if $\{f(\cdot;\theta_k)\}$ is chosen as the Gaussian family, then $E_{\theta_k}[T(X)]$ reduces to the mean and second moment of the Gaussian distribution.

Based on the updating scheme of θ, we propose the following algorithm for solving (1).

Algorithm 1 Gradient-Based Adaptive Stochastic Search (GASS)

1. *Initialization:* choose an exponential family of densities $\{f(\cdot;\theta)\}$, and specify a small positive constant ϵ, initial parameter θ_0, sample size sequence $\{N_k\}$, and step size sequence $\{\alpha_k\}$. Set $k = 0$.

2. *Sampling:* draw samples $x_k^i \overset{\text{iid}}{\sim} f(x;\theta_k), i = 1, 2, \ldots, N_k$.

3. *Estimation:* compute the normalized weights $\widehat{w}_k^i$ according to

$$\widehat{w}_k^i = \frac{\widehat{S}_{\theta_k}(H(x_k^i))}{\sum_{j=1}^{N_k} \widehat{S}_{\theta_k}(H(x_k^j))},$$

   and then compute $\widehat{E}_{p_k}[T(X)]$ and $\widehat{\text{Var}}_{\theta_k}[T(X)]$ respectively according to (10) and (11).

4. *Updating:* update the parameter θ according to

$$\theta_{k+1} = \Pi_{\tilde{\Theta}} \left\{ \theta_k + \alpha_k (\widehat{\text{Var}}_{\theta_k}[T(X)] + \epsilon I)^{-1} (\widehat{E}_{p_k}[T(X)] - E_{\theta_k}[T(X)]) \right\},$$

   where $\tilde{\Theta} \subseteq \Theta$ is a non-empty compact connected constraint set, and $\Pi_{\tilde{\Theta}}$ denotes the projection operator that projects an iterate back onto the set $\tilde{\Theta}$ by choosing the closest point in $\tilde{\Theta}$.

5. *Stopping:* check if some stopping criterion is satisfied. If yes, stop and return the current best sampled solution; else, set $k := k + 1$ and go back to step 2.

In the above algorithm, at the k^{th} iteration candidate solutions are drawn from the sampling distribution $f(\cdot;\theta_k)$, and then are used to estimate the quantities in the updating equation for θ_k so as to generate the next sampling distribution $f(\cdot;\theta_{k+1})$. To develop an intuitive understanding of the algorithm, we consider the special setting $T(X) = X$, in which case the term $\widehat{\text{Var}}_{\theta_k}[T(X)]$ basically measures how widespread the candidate solutions are. Since the magnitude of the ascent step is determined by $(\widehat{\text{Var}}_{\theta_k}[T(X)]+\epsilon I)^{-1}$, the algorithm takes smaller ascent steps to update θ when the candidate solutions are more widely spread (i.e., $\widehat{\text{Var}}_{\theta_k}[X]$ is larger), and takes larger ascent steps when the candidate solutions are more concentrated (i.e., $\widehat{\text{Var}}_{\theta_k}[X]$ is smaller). It means that exploitation of the local structure is adapted to our belief about the promising regions of the solution space: we will be more conservative in exploitation if we are uncertain about where the promising regions are and more progressive otherwise. Note that the projection operator at step 4 is primarily used to ensure the numerical stability of the algorithm. It prevents the iterates of the algorithm from becoming too big in practice and ensures the sequence

$\{\theta_k\}$ to stay bounded as the search proceeds. For simplicity, we will assume that $\tilde{\Theta}$ is a hyper-rectangle and takes the form $\tilde{\Theta} = \{\theta \in \Theta : a_i \leq \theta_i \leq b_i\}$ for constants $a_i < b_i$, $i = 1, \ldots, d$; other more general choices of $\tilde{\Theta}$ may also be used (see, e.g., Section 4.3 of [13]). Intuitively, such a constraint set should be chosen sufficiently large in practice so that the limits of the recursion at step 4 without the projection are contained in its interior.

3.1 Accelerated GASS

GASS can be viewed as a stochastic approximation (SA) algorithm, which we will show in more details in the next section. To improve the convergence rate of SA algorithms, [20] and [24] first proposed to take the average of the θ values generated by previous iterations, which is often referred to as Polyak (or Polyak-Ruppert) averaging. The original Polyak averaging technique is "offline", i.e., the averages are not fed back into the iterates of θ, and hence the averages are not useful for guiding the stochastic search in our context. However, there is a variation, Polyak averaging with online feedback (c.f. pp. 75 - 76 in [13]), which is not optimal as the original Polyak averaging but also enhances the convergence rate of SA. Using the Polyak averaging with online feedback, the parameter θ will be updated according to

$$\theta_{k+1} = \Pi_{\tilde{\Theta}} \left\{ \theta_k + \alpha_k \left(\widehat{\mathrm{Var}}_{\theta_k}[T(X)] + \epsilon I \right)^{-1} (\widehat{E}_{p_k}[T(X)] - E_{\theta_k}[T(X)]) + \alpha_k c(\bar{\theta}_k - \theta_k) \right\}, \tag{12}$$

where the constant c is the feedback weight, and $\bar{\theta}_k$ is the average

$$\bar{\theta}_k = \frac{1}{k} \sum_{i=1}^{k} \theta_i,$$

which can be calculated recursively by

$$\bar{\theta}_k = \frac{k-1}{k} \bar{\theta}_{k-1} + \frac{\theta_k}{k}. \tag{13}$$

With this parameter updating scheme, we propose the following algorithm.

Algorithm 2 Gradient-based Adaptive Stochastic Search with Averaging (GASS_avg)

Same as Algorithm 1 except in step 4 the parameter updating follows (12) and (13).

3.2 Derivation

In this subsection, we explain the rationale behind the updating scheme (8). We first derive the expressions of the gradient and Hessian of $L(\theta; \theta')$ as given below.

Proposition 1. *Assume that $f(x;\theta)$ is twice differentiable on Θ and that $\nabla_\theta f(x;\theta)$ and $\nabla^2_\theta f(x;\theta)$ are both bounded on $\mathcal{X}$ for any $\theta \in \Theta$. Then*

$$\begin{aligned}\nabla_\theta L(\theta;\theta') &= E_\theta[S_{\theta'}(H(X))\nabla_\theta \ln f(X;\theta)] \\ \nabla^2_\theta L(\theta;\theta') &= E_\theta[S_{\theta'}(H(X))\nabla^2_\theta \ln f(X;\theta)] \\ &\quad + E_\theta[S_{\theta'}(H(X))\nabla_\theta \ln f(X;\theta)\nabla_\theta \ln f(X;\theta)^T].\end{aligned}$$

Furthermore, if $f(x;\theta)$ is in an exponential family of densities defined by (6), then the above expressions reduce to

$$\begin{aligned}\nabla_\theta L(\theta;\theta') &= E_\theta[S_{\theta'}(H(X))T(X)] - E_\theta[S_{\theta'}(H(X))]E_\theta[T(X)], \\ \nabla^2_\theta L(\theta;\theta') &= E_\theta\left[S_{\theta'}(H(X))(T(X) - E_\theta[T(X)])(T(X) - E_\theta[T(X)])^T\right] \\ &\quad - \mathrm{Var}_\theta[T(X)]E_\theta[S_{\theta'}(H(X))].\end{aligned}$$

Notice that if we were to use Newton's method to update the parameter θ, the Hessian $\nabla^2_\theta L(\theta;\theta')$ is not necessarily negative semidefinite to ensure the parameter updating is along the ascent direction of $L(\theta;\theta')$, so we need some stabilization scheme. One way is to approximate the Hessian by the second term on the righthand side with a small perturbation, i.e., $-(\mathrm{Var}_\theta[T(X)] + \epsilon I)E_\theta[S_{\theta'}(H(X))]$, which is always negative definite. Thus, the parameter θ could be updated according to the following iteration

$$\begin{aligned}\theta_{k+1} &= \theta_k + \alpha_k \left((\mathrm{Var}_{\theta_k}[T(X)] + \epsilon I)E_{\theta_k}[S_{\theta_k}(H(X))]\right)^{-1} \nabla_\theta L(\theta_k;\theta_k), \\ &= \theta_k + \alpha_k \left(\mathrm{Var}_{\theta_k}[T(X)] + \epsilon I\right)^{-1} \left(\frac{E_{\theta_k}[S_{\theta_k}(H(X))T(X)]}{E_{\theta_k}[S_{\theta_k}(H(X))]} - E_{\theta_k}[T(X)]\right),\end{aligned} \tag{14}$$

which immediately leads to the updating scheme (8) given before.

In the updating equation (8), the term $E_{\theta_k}[S_{\theta_k}(H(X))]^{-1}$ is absorbed into $\nabla_\theta L(\theta_k;\theta_k)$, so we obtain a scale-free term $(E_{p_k}[T(X)] - E_{\theta_k}[T(X)])$ that is not subject to the scaling of the function value of $S_{\theta_k}(H(x))$. It would be nice to have such a scale-free gradient so that we can employ other gradient-based methods more easily besides the above specific choice of a quasi-Newton method. Towards this direction, we consider a further transformation of the maximization problem (5) by letting

$$l(\theta;\theta') = \ln L(\theta;\theta').$$

Since $\ln : R^+ \to R$ is a strictly increasing function, the maximization problem (5) is equivalent to

$$\theta^* = \arg\max_{\theta\in\mathbb{R}^d} l(\theta;\theta'). \tag{15}$$

The gradient and the Hessian of $l(\theta;\theta')$ are given in the following proposition.

Proposition 2. *Assume that $f(x;\theta)$ is twice differentiable on Θ and that $\nabla_\theta f(x;\theta)$ and $\nabla^2_\theta f(x;\theta)$ are both bounded on $\mathcal{X}$ for any $\theta \in \Theta$. Then*

$$\begin{aligned}\nabla_\theta l(\theta;\theta')|_{\theta=\theta'} &= E_{p(\cdot;\theta')}[\nabla_\theta \ln f(X;\theta')] \\ \nabla^2_\theta l(\theta;\theta')|_{\theta=\theta'} &= E_{p(\cdot;\theta')}[\nabla^2_\theta \ln f(X;\theta')] + \mathrm{Var}_{p(\cdot;\theta')}\left[\nabla_\theta \ln f(X;\theta')\right].\end{aligned}$$

Furthermore, if $f(x;\theta)$ is in an exponential family of densities, then the above expressions reduce to

$$\begin{aligned}\nabla_\theta l(\theta;\theta')|_{\theta=\theta'} &= E_{p(\cdot;\theta')}[T(X)] - E_{\theta'}[T(X)], \\ \nabla^2_\theta l(\theta;\theta')|_{\theta=\theta'} &= \mathrm{Var}_{p(\cdot;\theta')}[T(X)] - \mathrm{Var}_{\theta'}[T(X)].\end{aligned}$$

Similarly as before, noticing that the Hessian $\nabla^2_\theta l(\theta';\theta')$ is not necessarily negative definite to ensure the parameter updating is along the ascent direction of $l(\theta;\theta')$, we approximate the Hessian by the slightly perturbed second term in $\nabla^2_\theta l(\theta';\theta')$, i.e., $-(\mathrm{Var}_{\theta'}[T(X)] + \epsilon I)$. Then by setting

$$\theta_{k+1} = \theta_k + \alpha_k \left(\mathrm{Var}_{\theta_k}[T(X)] + \epsilon I\right)^{-1} \nabla_\theta l(\theta_k;\theta_k),$$

we again obtain exactly the same updating equation (8) for θ. The difference from (14) is that the gradient $\nabla_\theta l(\theta;\theta')$ is a scale-free term, and hence can be used in other gradient-based methods with easier choices of the step size. From the algorithmic viewpoint, it is better to consider the optimization problem (15) on $l(\theta;\theta')$ instead of the problem (5) on $L(\theta;\theta')$, even though both have the same global optima.

Although there are many ways to determine the positive definite matrix in front of the gradient in a quasi-Newton method, our choice of $(\mathrm{Var}_{\theta_k}[T(X)] + \epsilon I)^{-1}$ is not arbitrary but based on some principle. Without considering the numerical stability and thus dropping the term ϵI, the term $\mathrm{Var}_\theta[T(X)] = E[\nabla_\theta \ln f(X;\theta)(\nabla_\theta \ln f(X;\theta))^T] = E[-\nabla^2_\theta \ln f(X;\theta)]$ is the Fisher information matrix, whose inverse provides a lower bound on the variance of an unbiased estimator of the parameter θ ([21]), leading to the fact that $(\mathrm{Var}_\theta[T(X)])^{-1}$ is the minimum-variance step size in stochastic approximation ([16]). Moreover, from the optimization perspective, the term $(\mathrm{Var}_\theta[T(X)])^{-1}$ relates the gradient search on the parameter space with the stochastic search on the solution space, and thus adaptively adjusts the updating of the sampling distribution to our belief about the promising regions of the solution space. To see this more easily, let us consider $T(X) = X$. Then $(\mathrm{Var}_\theta[X])^{-1}$ is smaller (i.e., the gradient step in updating θ is smaller) when the current sampling distribution is more flat, signifying the exploration of the solution space is still active and we do not have a strong belief (i.e. $f(\cdot;\theta)$) about promising regions; $(\mathrm{Var}_\theta[X])^{-1}$ is larger (i.e., the gradient step in updating θ is larger) when our belief $f(\cdot;\theta)$ is more focused on some promising regions.

4. Convergence Analysis

We will analyze the convergence properties of GASS, resorting to methods and results in stochastic approximation (e.g., [12, 13, 1]). In GASS, $\nabla_\theta l(\theta;\theta_k)|_{\theta=\theta_k}$ is estimated by

$$\widehat{\nabla}_\theta l(\theta_k;\theta_k) = \widehat{E}_{p_k}[T(X)] - E_{\theta_k}[T(X)]. \tag{16}$$

To simplify notations, we denote

$$\widehat{V}_k \triangleq \widehat{\mathrm{Var}}_{\theta_k}[T(X)] + \epsilon I, \quad V_k \triangleq \mathrm{Var}_{\theta_k}[T(X)] + \epsilon I.$$

Hence, the parameter updating iteration in GASS is

$$\theta_{k+1} = \Pi_{\tilde{\Theta}}\left\{\theta_k + \alpha_k \widehat{V}_k^{-1} \widehat{\nabla}_\theta l(\theta_k;\theta_k)\right\}, \tag{17}$$

which can be rewritten in the form of a generalized Robbins-Monro algorithm

$$\theta_{k+1} = \theta_k + \alpha_k[D(\theta_k) + b_k + \xi_k + z_k], \tag{18}$$

where

$$\begin{aligned} D(\theta_k) &= \left(\mathrm{Var}_{\theta_k}[T(X)] + \epsilon I\right)^{-1} \nabla_\theta l(\theta_k;\theta_k), \\ b_k &= \widehat{V}_k^{-1}\left(\widehat{E}_{p_k}[T(X)] - \widetilde{E}_{p_k}[T(X)]\right), \\ \xi_k &= \left(\widehat{V}_k^{-1} - V_k^{-1}\right)\left(\widetilde{E}_{p_k}[T(X)] - E_{\theta_k}[T(X)]\right) + V_k^{-1}\left(\widetilde{E}_{p_k}[T(X)] - E_{p_k}[T(X)]\right), \end{aligned}$$

and z_k is the projection term satisfying $\alpha_k z_k = \theta_{k+1} - \theta_k - \alpha_k[D(\theta_k) + b_k + \xi_k]$, the minimum Euclidean length vector that takes the current iterate back onto the constraint set. The term $D(\theta_k)$ is the gradient vector field, b_k is the bias due to the inexact evaluation of the shape function in $\widehat{E}_{p_k}[T(X)]$ (b_k is zero if the shape function can be evaluated exactly), and ξ_k is the noise term due to Monte Carlo sampling in the approximations $\widehat{\mathrm{Var}}_{\theta_k}[T(X)]$ and $\widetilde{E}_{p_k}[T(X)]$.

For a given $\theta \in \tilde{\Theta}$, we define a set $C(\theta)$ as follows: if θ lies in the interior of $\tilde{\Theta}$, let $C(\theta) = \{0\}$; if θ lies on the boundary of $\tilde{\Theta}$, define $C(\theta)$ as the infinite convex cone generated by the outer normals at θ of the faces on which θ lies ([13] pp. 106). The difference equation (18) can be viewed as a noisy discretization of the constrained ordinary differential equation (ODE)

$$\dot{\theta}_t = D(\theta_t) + z_t, \quad z_t \in -C(\theta_t), \quad t \geq 0, \tag{19}$$

where z_t is the minimum force needed to keep the trajectory of the ODE in $\tilde{\Theta}$. Thus, the sequence of $\{\theta_k\}$ generated by (18) can be shown to asymptotically approach the solution set of the above ODE (19) by using the well-known ODE method. Let $\|\cdot\|$ denote the vector supremum norm (i.e., $\|x\| = \max\{|x_i|\}$) or the matrix max norm (i.e.,

$\|A\| = \max\{|a_{ij}|\}$). Let $\|\cdot\|_2$ denote the vector 2-norm (i.e., $\|x\| = \sqrt{x_1^2 + \ldots + x_n^2}$) or the matrix norm induced by the vector 2-norm (also called spectral norm for a square matrix, i.e., $\|A\|_2 = \sqrt{\lambda_{max}(A^*A)}$, where A^* is the conjugate transpose of A and λ_{max} returns the largest eigenvalue).

To proceed to the formal analysis, we introduce the following notations and assumptions. We denote the sequence of increasing sigma-fields generated by all the samples up to the k^{th} iteration by

$$\left\{\mathcal{F}_k = \sigma\left(\{x_0^i\}_{i=1}^{N_0}, \{x_1^i\}_{i=1}^{N_1}, \ldots, \{x_k^i\}_{i=1}^{N_k}\right), k = 0, 1, \ldots\right\}.$$

Define notations

$$\bar{\mathbb{U}}_k := \frac{1}{N_k}\sum_{i=1}^{N_k} \widehat{S}_{\theta_k}(H(x_k^i))T(x_k^i),\ \bar{\mathbb{V}}_k := \frac{1}{N_k}\sum_{i=1}^{N_k} \widehat{S}_{\theta_k}(H(x_k^i))$$
$$\tilde{\mathbb{U}}_k := \frac{1}{N_k}\sum_{i=1}^{N_k} S_{\theta_k}(H(x_k^i))T(x_k^i),\ \tilde{\mathbb{V}}_k := \frac{1}{N_k}\sum_{i=1}^{N_k} S_{\theta_k}(H(x_k^i))$$
$$\mathbb{U}_k := E_{\theta_k}[S_{\theta_k}(H(X))T(X)],\ \mathbb{V}_k := E_{\theta_k}[S_{\theta_k}(H(X))].$$

Assumption 1.
(i) The step size sequence $\{\alpha_k\}$ satisfies $\alpha_k > 0$ for all k, $\alpha_k \searrow 0$ as $k \to \infty$, and $\sum_{k=0}^{\infty}\alpha_k = \infty$.
(ii) The sample size $N_k = N_0 k^{\zeta}$, where $\zeta > 0$; moreover, $\{\alpha_k\}$ and $\{N_k\}$ jointly satisfies $\frac{\alpha_k}{\sqrt{N_k}} = O(k^{-\beta})$ for some constant $\beta > 1$.
(iii) The function $x \mapsto T(x)$ is bounded on $\mathcal{X}$.
(iv) For any x, $|\widehat{S}_{\theta_k}(H(x)) - S_{\theta_k}(H(x))| \to 0$ w.p.1 as $N_k \to \infty$.

In the above assumption, (i) is a typical assumption on the step size sequence in SA, which means that α_k diminishes not too fast. Assumption 1(ii) provides a guideline on how to choose the sample size given a choice of the step size sequence, and shows that the sample size has to increase to infinity no slower than a certain speed. For example, if we choose $\alpha_k = \alpha_0 k^{-\alpha}$ with $0 < \alpha < 1$, then it is sufficient to choose $N_k = O(k^{2(\beta-\alpha)})$. Assumption 1(iii) holds true for many exponential families used in practice. Assumption 1(iv) is a sufficient condition to ensure the strong consistency of estimates, and is satisfied by many choices of the shape function S_θ. For example, it is trivially satisfied if $S_\theta = S$, since $S(H(x))$ can be evaluated exactly for each x. If S_θ takes the form of (4), Assumption 1(iv) is also satisfied, as shown in the following lemma.

Lemma 1. *Suppose the shape function takes the form*

$$S_{\theta_k}(H(x)) = (H(x) - H_{lb})\frac{1}{1 + e^{S_0(H(x) - \gamma_{\theta_k})}},$$

where $\gamma_{\theta_k} \triangleq \sup_l \{l : P_{\theta_k}\{x \in \mathcal{X} : H(x) \geq l\} \geq \rho\}$ *is the unique* $(1-\rho)$*-quantile with respect to* $f(\cdot;\theta_k)$*. If* $S_{\theta_k}(H(x))$ *is estimated by* $\widehat{S}_{\theta_k}(H(x))$ *with the true quantile* γ_{θ_k} *being replaced by the sample* $(1-\rho)$*-quantile* $\widehat{\gamma}_{\theta_k} = H_{(\lceil (1-\rho)N_k \rceil)}$*, where* $\lceil a \rceil$ *is the smallest integer greater than* a*, and* $H_{(i)}$ *is the* i^{th} *order statistic of the sequence* $\{H(x_k^i), i = 1, \ldots, N_k\}$*. Then under the condition* $N_k = \Theta(k^\zeta)$ $\zeta > 0$*, we have that for every* x*,* $\left|\widehat{S}_{\theta_k}(H(x)) - S_{\theta_k}(H(x))\right| \to 0$ *w.p.*1 as $k \to \infty$.

The next lemma shows that the summed tail error goes to zero w.p.1.

Lemma 2. *Under Assumption 1 (i)-(iii), for any* $T > 0$*,*

$$\lim_{k\to\infty} \left\{ \sup_{\{n:0\leq\sum_{i=k}^{n-1}\alpha_i\leq T\}} \left\| \sum_{i=k}^{n} \alpha_i \xi_i \right\| \right\} = 0, \quad w.p.1.$$

Theorem 1 below shows that GASS generates a sequence $\{\theta_k\}$ that asymptotically approaches the limiting solution of the ODE (19) under the regularity conditions specified in Assumption 1.

Theorem 1. *Assume that* $D(\theta_t)$ *is continuous with a unique integral curve (i.e., the ODE (19) has a unique solution* $\theta(t)$*) and Assumption 1 holds. Then the sequence* $\{\theta_k\}$ *generated by (17) converges to a limit set of (19) w.p.1. Furthermore, if the limit sets of (19) are isolated equilibrium points, then w.p.1* $\{\theta_k\}$ *converges to a unique equilibrium point.*

For a given distribution family, Theorem 1 shows that our algorithm will identify a local/global optimal sampling distribution within the given family that provides the best capability in generating an optimal solution to (1). From the viewpoint of maximizing $E_\theta[H(X)]$, the average function value under our belief of where promising solutions are located (i.e., the parameterized distribution $f(x,\theta)$), the convergence of the algorithm to an local/global optimum in the parameter space essentially gives us a local/global optimum of our belief about the function value.

4.1 Asymptotic Normality of GASS

In this section, we study the asymptotic convergence rate of Algorithm 1 under the assumption that the parameter sequence $\{\theta_k\}$ converges to a unique equilibrium point θ^* of the ODE (19) in the interior of $\tilde{\Theta}$. This indicates that there exists a small open neighborhood $\mathcal{N}(\theta^*)$ of θ^* such that the sequence $\{\theta_k\}$ will be contained in $\mathcal{N}(\theta^*)$ for k sufficiently large w.p.1. Thus, the projection operator in (17) and z_k in (18) can be dropped in the analysis, because the projected recursion will behave identically to an unconstrained algorithm in the long run. Define $\mathcal{L}(\theta) = \nabla_{\theta'} l(\theta';\theta)|_{\theta'=\theta}$ and let $J_{\mathcal{L}}$ be the Jacobian of $\mathcal{L}$. Under our conditions, it immediately follows from (19) that $C(\theta^*) = \{0\}$ and $\mathcal{L}(\theta^*) = 0$. Since $\mathcal{L}$ is the gradient of some underlying function $F(\theta)$, $J_{\mathcal{L}}$ is the Hessian of F and Algorithm 1 is essentially a gradient-based algorithm for maximizing $F(\theta)$. Therefore, it is reasonable to expect that the following assumption holds:

Assumption 2. *The Hessian matrix $J_{\mathcal{L}}(\theta)$ is continuous and symmetric negative definite in the neighborhood $\mathcal{N}(\theta^*)$ of θ^*.*

We consider a standard gain sequence $\alpha_k = \alpha_0/k^\alpha$ for constants $\alpha_0 > 0$ and $0 < \alpha < 1$, a polynomially increasing sample size $N_k = N_0 k^\zeta$ with $N_0 \geq 1$ and $\zeta > 0$.

By dropping the projection operator in (17), we can rewrite the equation in the form:

$$\delta_{k+1} = \delta_k + k^{-\alpha}\Phi_k \mathcal{L}(\theta_k) + k^{-\alpha}\Phi_k\Big(\frac{\bar{\mathbb{U}}_k}{\bar{\mathbb{V}}_k} - \frac{\mathbb{U}_k}{\mathbb{V}_k}\Big),$$

where $\delta_k = \theta_k - \theta^*$ and $\Phi_k = \alpha_0(\widehat{\mathrm{Var}}_{\theta_k}(T(X)) + \epsilon I)^{-1}$. Next, by using a first order Taylor expansion of $\mathcal{L}(\theta_k)$ around the neighborhood of θ^* and the fact that $\mathcal{L}(\theta^*) = 0$, we have

$$\delta_{k+1} = \delta_k + k^{-\alpha}\Phi_k J_{\mathcal{L}}(\tilde{\theta}_k)\delta_k + k^{-\alpha}\Phi_k\Big(\frac{\bar{\mathbb{U}}_k}{\bar{\mathbb{V}}_k} - \frac{\mathbb{U}_k}{\mathbb{V}_k}\Big),$$

where $\tilde{\theta}_k$ lies on the line segment from θ_k to θ^*. For a given positive constant $\tau > 0$, the above equation can be further written in the form of a recursion in [5]:

$$\delta_{k+1} = (I - k^{-\alpha}\Gamma_k)\delta_k + k^{-(\alpha+\tau)/2}\Phi_k W_k + k^{-\alpha-\tau/2}T_k,$$

where $\Gamma_k = -\Phi_k J_{\mathcal{L}}(\tilde{\theta}_k)$, $W_k = k^{(\tau-\alpha)/2}\big(\frac{\bar{\mathbb{U}}_k}{\bar{\mathbb{V}}_k} - E_{\theta_k}\big[\frac{\bar{\mathbb{U}}_k}{\bar{\mathbb{V}}_k}\big|\mathcal{F}_{k-1}\big]\big)$, and $T_k = k^{\tau/2}\Phi_k\big(\frac{\bar{\mathbb{U}}_k}{\bar{\mathbb{V}}_k} - \frac{\tilde{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k} + E_{\theta_k}\big[\frac{\tilde{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k}\big|\mathcal{F}_{k-1}\big] - \frac{\mathbb{U}_k}{\mathbb{V}_k}\big)$. The basic idea of the rate analysis is to show that the sequence of amplified differences $\{k^{\tau/2}\delta_k\}$ converges in distribution to a normal random variable with mean zero and constant covariance matrix. To this end, we show that all sufficient conditions in Theorem 2.2 in [5] are satisfied in our setting. We begin with a strengthened version of Assumption 1(*iv*).

Assumption 3.
For a given constant $\tau > 0$ and $x \in \mathcal{X}$, $k^{\tau/2}|\widehat{S}_{\theta_k}(H(x)) - S_{\theta_k}(H(x))| \to 0$ as $k \to \infty$ w.p.1.

Assumption 3 holds trivially when S_θ is a deterministic function that is independent of θ. In addition, if sample quantiles are involved in the shape function and $S_{\theta_k}(H(x))$ takes the form (4), then the assumption can also be justified under some additional mild regularity conditions; cf. e.g., [9].

Let $\Phi = \alpha_0(\mathrm{Var}_{\theta^*}(T(X)) + \epsilon I)^{-1}$ and $\Gamma = -\Phi J_{\mathcal{L}}(\theta^*)$. The following result shows condition (2.2.1) in Theorem 2.2 of [5].

Lemma 3. *Assume Assumptions 1 and 2 hold, we have $\Phi_k \to \Phi$ and $\Gamma_k \to \Gamma$ as $k \to \infty$ w.p.1. In addition, if Assumption 1(iv) is replaced with Assumption 3 and $N_k = N_0 k^\zeta$ with $\zeta > \tau/2$, then $T_k \to 0$ as $k \to \infty$ w.p.1.*

In addition, the noise term W_k has the following property, which justifies condition 2.2.2 in [5].

Lemma 4. *$E_{\theta_k}[W_k|\mathcal{F}_{k-1}] = 0$. In addition, let τ be a given constant satisfying $\tau > \alpha$. If Assumption 1 holds and $N_k = N_0 k^{\tau-\alpha}$, then there exists a positive semi-definite matrix Σ such that $\lim_{k\to\infty} E_{\theta_k}[W_k W_k^T|\mathcal{F}_{k-1}] = \Sigma$ w.p.1, and $\lim_{k\to\infty} E[I\{\|W_k\|^2 \geq rk^\alpha\}\|W_k\|^2] = 0\ \forall r > 0$.*

The following asymptotic normality results then follows directly from Theorem 2.2 in [5].

Theorem 2. *Let $\alpha_k = \alpha_0/k^\alpha$ for $0 < \alpha < 1$. For a given constant $\tau > 2\alpha$, let $N_k = N_0 k^{\tau-\alpha})$. Assume the convergence of the sequence $\{\theta_k\}$ occurs to a unique equilibrium point θ^* w.p.1. If Assumptions 1, 2, and 3 hold, then*

$$k^{\frac{\tau}{2}}(\theta_k - \theta^*) \xrightarrow{dist} N(0, Q\mathcal{M}Q^T),$$

where Q is an orthogonal matrix such that $Q^T(-J_{\mathcal{L}}(\theta^*))Q = \Lambda$ with Λ being a diagonal matrix, and the $(i,j)^{th}$ entry of the matrix $\mathcal{M}$ is given by $\mathcal{M}_{(i,j)} = (Q^T\Phi\Sigma\Phi^T Q)_{(i,j)}(\Lambda_{(i,i)} + \Lambda_{(j,j)})^{-1}$.

Theorem 2 shows the asymptotic rate at which the noise caused by Monte-Carlo random sampling in GASS will be damped out as the number of iterations $k \to \infty$. This rate, as indicated in the theorem, is on the order of $O(1/\sqrt{k^\tau})$. This implies that the noise can be damped out arbitrarily fast by using a sample size sequence $\{N_k\}$ that increases sufficiently fast as $k \to \infty$. However, we note that this rate result is stated in terms of the number of iterations k, not the sample size N_k. Therefore, in practice, there is the need to carefully balance the tradeoff between the choice of large values of N_k to increase the algorithms's asymptotic rate and the use of small values of N_k to reduce the per iteration computational cost.

5. Numerical Experiments

We test the proposed algorithms GASS, GASS_avg on some benchmark continuous optimization problems selected from [8] and [9]. To fit in the maximization framework where our algorithms are proposed, we take the negative of those objective functions that are originally for minimization problems. The ten benchmark problems are listed as below.

(1) Dejong's 5th function (n=2, $-50 \leq x_i \leq 50$)

$$H_1(x) = -\left[0.002 + \sum_{j=1}^{25} \frac{1}{j + \sum_{i=1}^{2}(x_i - a_{ji})^6}\right]^{-1},$$

where $a_{j1} = (-32, -16, 0, 16, 32, -32, -16, 0, 16, 32, -32, -16, 0, 16, 32, -32, -16, 0, 16, 32, -32, -16, 0, 16, 32)$ and $a_{j2} = (-32, -32, -32, -32, -32, -16, -16, -16, -16, -16, 0, 0, 0, 0, 0, 16, 16, 16, 16, 16, 32, 32, 32, 32, 32)$. The global optimum is at $x^* = (-32, -32)^T$, and $H^* \approx -0.998$.

(2) Shekel's function (n=4, $0 \le x_i \le 10$)

$$H_2(x) = \sum_{i=1}^{5} \left((x-a_i)^T(x-a_i) + c_i\right)^{-1},$$

where $a_1 = (4,4,4,4)^T$, $a_2 = (1,1,1,1)^T$, $a_3 = (8,8,8,8)^T$, $a_4 = (6,6,6,6)^T$, $a_5 = (3,7,3,7)^T$, and $c = (0.1, 0.2, 0.2, 0.4, 0.4)$. $x^* = (4,4,4,4)^T$, $H^* \approx 10.153$.

(3) Powel singular function (n=50, $-50 \le x_i \le 50$)

$$H_3(x) = -\sum_{i=2}^{n-2} \left[(x_{i-1} + 10x_i)^2 + 5(x_{i+1} - x_{i+2})^2 + (x_i - 2x_{i+1})^4 + 10(x_{i-1} - x_{i+2})^4\right] - 1,$$

where $x^* = (0, \cdots, 0)^T$, $H^* = -1$.

(4) Rosenbrock function (n=10, $-10 \le x_i \le 10$)

$$H_4(x) = -\sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right] - 1,$$

where $x^* = (1, \cdots, 1)^T$, $H^* = -1$.

(5) Griewank function (n=50, $-50 \le x_i \le 50$)

$$H_5(x) = -\frac{1}{4000}\sum_{i=1}^{n} x_i^2 + \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) - 1,$$

where $x^* = (0, \cdots, 0)^T$, $H^* = 0$.

(6) Trigonometric function (n=50, $-50 \le x_i \le 50$)

$$H_6(x) = -\sum_{i=1}^{n} \left[8\sin^2(7(x_i - 0.9)^2) + 6\sin^2(14(x_i - 0.9)^2) + (x_i - 0.9)^2\right] - 1,$$

where $x^* = (0.9, \cdots, 0.9)^T$, $H^* = -1$.

(7) Rastrigin function (n=20, $-5.12 \le x_i \le 5.12$)

$$H_7(x) = -\sum_{i=1}^{n} \left(x_i^2 - 10\cos(2\pi x_i)\right) - 10n - 1,$$

where $x^* = (0, \cdots, 0)^T$, $H^* = -1$.

(8) Pintér's function (n=50, $-50 \le x_i \le 50$)

$$\begin{aligned} H_8(x) &= -\Bigg[\sum_{i=1}^{n} i x_i^2 + \sum_{i=1}^{n} 20i \sin^2(x_{i-1}\sin x_i - x_i + \sin x_{i+1}) \\ &\quad + \sum_{i=1}^{n} i \log_{10}(1 + i(x_{i-1}^2 - 2x_i + 3x_{i+1} - \cos x_i + 1)^2)\Bigg] - 1, \end{aligned}$$

where $x^* = (0, \cdots, 0)^T$, $H^* = -1$.

(9) Levy function (n=50, $-50 \le x_i \le 50$)

$$H_9(x) = -\sin^2(\pi y_1) - \sum_{i=1}^{n-1} \left[(y_i - 1)^2(1 + 10\sin^2(\pi y_i + 1))\right] - (y_n - 1)^2(1 + 10\sin^2(2\pi y_n)) - 1,$$

where $y_i = 1 + (x_i - 1)/4$, $x^* = (1, \cdots, 1)^T$, $H^* = -1$.

(10) Weighted Sphere function (n=50, $-50 \le x_i \le 50$)

$$H_{10}(x) = -\sum_{i=1}^{n} i x_i^2 - 1$$

where $x^* = (0, \cdots, 0)^T$, $H^* = -1$.

Specifically, Dejong's 5th (H_1) and Shekel's (H_2) are low-dimensional problems with a small number of local optima that are scattered and far from each other; Powel (H_3) and Rosenbrock (H_4) are badly-scaled functions; Griewank (H_5), Trigonometric (H_6), and Rastrigin (H_7) are high-dimensional multimodal problems with a large number of local optima, and the number of local optima increases exponentially with the problem dimension; Pintér (H_8) and Levy (H_9) are both multimodal and badly-scaled problems; Weighted Sphere function (H_{10}) is a high-dimensional concave function.

We compare the performance of GASS and GASS_avg with two other algorithms: the modified version of the CE method based on stochastic approximation proposed by [9] and the MRAS method proposed by [8]. In our comparison, we try to use the same parameter setting in all four methods. The common parameters in all four methods are set as follows: the quantile parameter is set to be $\rho = 0.02$ for low-dimensional problems H_1 and H_2, and $\rho = 0.05$ for all the other problems; the parameterized exponential family distribution $f(x; \theta_k)$ is chosen to be independent multivariate normal distribution $\mathcal{N}(\mu_k, \Sigma_k)$; the initial mean μ_0 is chosen randomly according to the uniform distribution on $[-30, 30]^n$, and the initial covariance matrix is set to be $\Sigma_0 = 1000 I_{n\times n}$, where n is the dimension of the problem; the sample size at each iteration is set to be $N = 1000$. In addition, we observe that the performance of the algorithm is insensitive to the initial candidate solutions if the initial variance is large enough.

In GASS and GASS_avg, we consider the shape function of the form (4), i.e.,

$$S_{\theta_k}(H(x)) = (H(x) - H_{lb}) \frac{1}{1 + e^{-S_0(H(x)-\gamma_{\theta_k})}},$$

In our experiment, S_0 is set to be 10^5, which makes $S_{\theta_k}(H(x))$ a very close approximation to $(H(x) - H_{lb})I\{H(x) \geq \gamma_{\theta_k}\}$; the $(1-\rho)$-quantile γ_{θ_k} is estimated by the $(1-\rho)$ sample quantile of the function values corresponding to all the candidate solutions generated at the k^{th} iteration. We use the step size: $\alpha_k = \alpha_0/k^\alpha$, where α_0 reflects the initial step size, and the parameter α should be between 0 and 1. We set $\alpha_0 = 0.3$ for the low-dimensional problems H_1 and H_2 and the badly-scaled problem H_4, and set $\alpha_0 = 1$ for the rest of the problems; we set $\alpha = 0.05$, which is chosen to be relatively small to provide a slowly decaying step size. With the above setting of step size, we can always find a β such that the sample size $N_k = 1000$ satisfies the Assumption 1(ii) under a finite number of iterations, e.g. $k < 2500$ in our experiment. In GASS_avg, the feedback weight is $c = 0.002$ for problems H_3, H_4 and H_8 and $c = 0.1$ for all other problems.

In the modified CE method, we use the gain sequence $\alpha_k = 5/(k + 100)^{0.501}$, which is found to work best in the experiments. In the implementation of MRAS method, we use a smoothing parameter ν when updating the parameter θ_k of the parameterized distribution, and set $\nu = 0.2$ as suggested by [8]. The rest of the parameter setting for MRAS is as follows: $\lambda = 0.01$, $r = 10^{-4}$ in the shape function $S(H(x)) = \exp\{rH(x)\}$. Other than using an increasing sample size in [9] and [8], and updating quantile ρ_k in [8], the constant sample size $N = 1000$ and a constant ρ are used in our experiments for a fair comparison of all the methods.

		GASS		GASS_avg		modified CE		MRAS	
	H^*	$\bar{H}^*(std\_err)$	M_ε	$\bar{H}^*(std\_err)$	M_ε	$\bar{H}^*(std\_err)$	M_ε	$\bar{H}^*(std\_err)$	M_ε
Dejong's 5th H_1	-0.998	-0.998(4.79E-7)	100	-0.998(8.97E-7)	100	-1.02(0.014)	95	-0.9981(6.63E-4)	98
Shekel H_2	10.153	9.92(0.114)	96	9.91(0.106)	95	10.153(1.09E-7)	79	9.90(0.126)	96
Powel H_3	-1	-1(1.48E-6)	100	-1(1.89E-6)	100	-1(8.87E-9)	100	-1.50(0.433)	95
Rosenbrock H_4	-1	-1.03(1.40E-4)	0	-1.09(0.0301)	46	-1.91(0.016)	0	-7.10(0.629)	0
Griewank H_5	0	0(8.45E-15)	100	0(7.30E-15)	100	-0(3.02E-16)	100	-0.14(0.017)	57
Trigonometric H_6	-1	-1(9.72E-13)	100	-1(1.08E-12)	100	-1(2.23E-18)	100	-1(4.69E-7)	100
Rastrigin H_7	-1	-1.15(0.0357)	85	-1.19(0.044)	83	-1.01(0.0099)	99	-83.45(0.634)	0
Pinter H_8	-1	-1.007(0.0034)	93	-1.04(0.0104)	63	-6.08(0.0254)	0	-530.4(48.64)	2
Levy H_9	-1	-1(9.56E-13)	100	-1(1.29E-7)	100	-1.063(3.87E-18)	100	-1(1.42E-10)	100
Sphere H_{10}	-1	-1(1.79E-11)	100	-1(1.42E-11)	100	-1(2.23E-18)	100	-1(9.95E-9)	100

Table 1: Comparison of GASS, GASS_avg, modified CE and MRAS

In the experiments, we found the computation time of function evaluations dominates the time of other steps, so we compare the performance of the algorithms with respect to the total number of function evaluations, which is equal to the total number of samples. The average performance based on 100 independent runs for each method is shown in Table

1, where H^* is the true optimal value of $H(\cdot)$; $\bar{H}^*$ is the average of the function values returned by the 100 runs of an algorithm; *std_err* is the standard error of these 100 function values; M_ε is the number of ε-optimal solutions out of 100 runs (ε-optimal solution is the solution such that $H^* - \hat{H}^* \leq \varepsilon$, where $\hat{H}^*$ is the optimal function value returned by an algorithm). We consider $\varepsilon = 10^{-2}$ for problems H_4, H_7, H_8 and $\varepsilon = 10^{-3}$ for all other problems. Fig. 1 and Fig. 2 show the average (over 100 runs) of best value of $H(\cdot)$ at the current iteration versus the total number of samples generated so far.

From the results, GASS and GASS_avg find all the ε-optimal solutions in 100 runs for problems H_1, H_3, H_5, H_6, H_9, and H_{10}. Modified CE finds all the ε-optimal solutions for problems H_3, H_5, H_6, H_9, and H_{10}. MRAS only finds all the ε-optimal solutions for the problems H_6 and H_9 and the convex problem H_{10}. As for the convergence rate, GASS_avg always converges faster than GASS, verifying the effectiveness of averaging with online feedback. Both GASS and GASS_avg converge faster than MRAS on all the problems, and converge faster than the modified CE method when α_0 is set to be large, i.e. on problems H_3 and $H_5 - H_{10}$.

6. Conclusion

In this paper, we have introduced a new model-based stochastic search algorithm for solving general black-box optimization problems. The algorithm generates candidate solutions from a parameterized sampling distribution over the feasible region, and uses a quasi-Newton like iteration on the parameter space of the parameterized distribution to find improved sampling distributions. Thus, the algorithm enjoys the fast convergence speed of classical gradient search methods while retaining the robustness feature of model-based methods. By formulating the algorithm iteration into the form of a generalized stochastic approximation recursion, we have established the convergence and convergence rate results of the algorithm. Our numerical results indicate that the algorithm shows promising performance as compared with some of the existing approaches.

A. Appendix

Proof. **Proof of Proposition 1.** Consider the gradient of $L(\theta;\theta')$ with respect to θ,

$$\begin{aligned} \nabla_\theta L(\theta;\theta') &= \int S_{\theta'}(H(x))\nabla_\theta f(x;\theta)dx \\ &= \int S_{\theta'}(H(x))f(x;\theta)\nabla_\theta \ln f(x;\theta)dx \\ &= E_\theta[S_{\theta'}(H(X))\nabla_\theta \ln f(X;\theta)], \end{aligned} \tag{20}$$

where the interchange of integral and derivative in the first equality follows from the boundedness assumptions on $S_{\theta'}$ and $\nabla_\theta f(x;\theta)$ and the dominated convergence theorem.

Consider the Hessian of $L(\theta;\theta')$ with respect to θ,

$$\begin{aligned}\nabla^2_\theta L(\theta;\theta') &= \int S_{\theta'}(H(x))\nabla^2_\theta f(x;\theta)dx \\ &= \int S_{\theta'}(H(x))f(x;\theta)\nabla^2_\theta \ln f(x;\theta)dx + \int S_{\theta'}(H(x))\nabla_\theta \ln f(x;\theta)\nabla_\theta f(x;\theta)^T dx \\ &= E_\theta[S_{\theta'}(H(X))\nabla^2_\theta \ln f(X;\theta)] + E_\theta[S_{\theta'}(H(X))\nabla_\theta \ln f(x;\theta)\nabla_\theta \ln f(x;\theta)^T] \quad (21)\end{aligned}$$

where the last equality follows from the fact that $\nabla_\theta f(x;\theta) = f(x;\theta)\nabla_\theta \ln f(x;\theta)$.

Furthermore, if $f(x;\theta) = \exp\{\theta^T T(x) - \phi(\theta)\}$, we have

$$\begin{aligned}\nabla_\theta \ln f(x;\theta) &= \nabla_\theta \left(\theta^T T(x) - \ln \int \exp(\theta^T T(x))dx\right) \\ &= T(x) - \frac{\int \exp(\theta^T T(x))T(x)dx}{\int \exp(\theta^T T(x))dx} \\ &= T(x) - E_\theta[T(X)]. \quad (22)\end{aligned}$$

Plugging (22) into (20) yields

$$\nabla_\theta L(\theta;\theta') = E_\theta[S_{\theta'}(H(X))T(X)] - E_\theta[S_{\theta'}(H(X))]E_\theta[T(X)].$$

Differentiating (22) with respect to θ, we obtain

$$\begin{aligned}\nabla^2_\theta \ln f(x;\theta) &= -\frac{\int \exp(\theta^T T(x))T(x)T(x)^T dx}{\int \exp(\theta^T T(x))dx} \\ &\quad + \frac{\int \exp(\theta^T T(x))T(x)dx \left(\int \exp(\theta^T T(x))T(x)dx\right)^T}{\left(\int \exp(\theta^T T(x))dx\right)^2} \\ &= -E_\theta[T(X)T(X)^T] + E_\theta[T(X)]E_\theta[T(X)]^T \\ &= -\mathrm{Var}_\theta[T(X)]. \quad (23)\end{aligned}$$

Plugging (22) and (23) into (21) yields

$$\begin{aligned}\nabla^2_\theta L(\theta;\theta') &= E_\theta[S_{\theta'}(H(X))(T(X) - E_\theta[T(X)])(T(X) - E_\theta[T(X)])^T] \\ &\quad - \mathrm{Var}_\theta[T(X)]E_\theta[S_{\theta'}(H(X))].\end{aligned}$$

□

Proof. **Proof of Proposition 2.** Consider the gradient of $l(\theta;\theta')$ with respect to θ,

$$\begin{aligned}\nabla_\theta l(\theta;\theta')|_{\theta=\theta'} &= \left.\frac{\nabla_\theta L(\theta;\theta')}{L(\theta;\theta')}\right|_{\theta=\theta'} \\ &= \left.\frac{\int S_{\theta'}(H(x))f(x;\theta)\nabla_\theta \ln f(x;\theta)dx}{L(\theta;\theta')}\right|_{\theta=\theta'} \quad (24) \\ &= E_{p(\cdot;\theta')}[\nabla_\theta \ln f(X;\theta')].\end{aligned}$$

Differentiating (24) with respect to θ, we obtain the Hessian

$$\begin{aligned}\nabla_\theta^2 l(\theta;\theta')|_{\theta=\theta'} &= \frac{\int S_{\theta'}(H(x))f(x;\theta)\nabla_\theta^2 \ln f(x;\theta)dx}{L(\theta;\theta')} + \frac{\int S_{\theta'}(H(x))\nabla_\theta \ln f(x;\theta)(\nabla_\theta f(x;\theta))^T dx}{L(\theta;\theta')}... \\ &\quad - \frac{(\int S_{\theta'}(H(x))f(x;\theta)\nabla_\theta \ln f(x;\theta)dx)(\nabla_\theta L(\theta;\theta'))^T}{L(\theta;\theta')^2}\Big|_{\theta=\theta'}\end{aligned}$$

Using $\nabla_\theta f(x;\theta) = f(x;\theta)\nabla_\theta \ln f(x;\theta)$ in the second term on the righthand side, the above expression can be written as

$$\begin{aligned}\nabla_\theta^2 l(\theta;\theta')|_{\theta=\theta'} &= E_{p(\cdot;\theta')}[\nabla_\theta^2 \ln f(X;\theta')] + E_{p(\cdot;\theta')}\left[\nabla_{\theta'} \ln f(X;\theta')(\nabla_{\theta'} \ln f(X;\theta'))^T\right] \\ &\quad - E_{p(\cdot;\theta')}\left[\nabla_\theta \ln f(X;\theta')\right] E_{p(\cdot;\theta')}\left[\nabla_\theta \ln f(X;\theta')\right]^T \\ &= E_{p(\cdot;\theta')}[\nabla_\theta^2 \ln f(X;\theta')] + \mathrm{Var}_{p(\cdot;\theta')}\left[\nabla_\theta \ln f(X;\theta')\right]. \qquad (25)\end{aligned}$$

Furthermore, if $f(x;\theta) = \exp\{\theta^T T(x) - \phi(\theta)\}$, plugging (22) into (24) yields

$$\nabla_\theta l(\theta;\theta')|_{\theta=\theta'} = E_{p(\cdot;\theta')}[T(X)] - E_{\theta'}[T(X)],$$

and plugging (22) and (23) into (25) yields

$$\nabla_\theta^2 l(\theta;\theta')|_{\theta=\theta'} = \mathrm{Var}_{p(\cdot;\theta')}[T(X)] - \mathrm{Var}_{\theta'}[T(X)].$$

□

Proof. **Proof of Lemma 1.** Because S_θ is continuous in γ_θ, it is sufficient to show that $\widehat{\gamma}_{\theta_k} \to \gamma_{\theta_k}$ w.p.1 as $k \to \infty$, which can be shown in the same way as Lemma 7 in [8], except that we need to verify the following condition in their proof:

$$\sum_{k=1}^{\infty} \exp\left(-\tilde{M} N_k\right) < \infty,$$

where $\tilde{M}$ is positive constant. It is easy to see that this condition is trivially satisfied in our setting by taking $N_k = N_0 k^\zeta$ with $\zeta > 0$. □

Proof. **Proof of Lemma 2.**

Before the formal proof of Lemma 2, we first introduce a key inequality to our proof - the matrix bounded differences inequality ([26]), which is a matrix version of the generalized Hoeffding inequality (i.e., McDiarmid's inequality ([15])). Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ return the largest and smallest eigenvalue of a matrix, respectively.

Theorem 3. *(Matrix bounded differences, Corollary 7.5, [26]) Let $\{X^i : i = 1, 2, \ldots, N\}$ be an independent family of random variables, and let V be a function that maps N variables to a self-adjoint matrix of dimension d. Consider a sequence of $\{C_k\}$ of fixed self-adjoint matrices that satisfy*

$$\left(V(x^1, \ldots, x^i, \ldots, x^N) - V(x^1, \ldots, \tilde{x}^i, \ldots, x^N)\right)^2 \leq C_i^2,$$

where x^i and $\tilde{x}^i$ range over all possible values of X^i for each index i. Compute the variance parameter

$$\sigma^2 := \left\| \sum_i C_i^2 \right\|_2.$$

Then, for all $\delta > 0$,

$$P\left\{\lambda_{\max}(V(\mathbf{x}) - E[V(\mathbf{x})]) \geq \delta\right\} \leq d \exp\left\{\frac{-\delta^2}{8\sigma^2}\right\},$$

where $\mathbf{x} = (X^1, \ldots, X^N)$.

Now we proceed to the formal proof of Lemma 2. Recall that ξ_k can be written as

$$\xi_k = (\widehat{V}_k^{-1} - V_k^{-1})(\widetilde{E}_{p_k}[T(X)] - E_{\theta_k}[T(X)]) + V_k^{-1}(\widetilde{E}_{p_k}[T(X)] - E_{p_k}[T(X)]). \tag{26}$$

To bound the first term on the right-hand-side in (26), we notice that since V_k^{-1} and $\widehat{V}_k^{-1}$ are both positive definite and $(\epsilon^{-1}I - V_k^{-1})$ and $(\epsilon^{-1}I - \widehat{V}_k^{-1})$ are both positive semi-definite, we have

$$\begin{aligned} \|V_k^{-1} - \widehat{V_k}^{-1}\| &= \|V_k^{-1}(\widehat{V}_k - V_k)\widehat{V}_k^{-1}\| \\ &\leq \|V_k^{-1}\|\|\widehat{V}_k - V_k\|\|\widehat{V}_k^{-1}\| \\ &\leq \epsilon^{-2}\|\widehat{V}_k - V_k\|. \end{aligned} \tag{27}$$

To establish a bound on $\|\widehat{V}_k - V_k\|$, we use the matrix bounded differences inequality that is introduced above. For simplicity of exposition, we drop the subscript k in the expression below.

$$\begin{aligned} &\sup_{x^i, \tilde{x}^i \in \mathcal{X}} \left\{\widehat{V}(x^1, \ldots, x^i, \ldots, x^N) - \widehat{V}(x^1, \ldots, \tilde{x}^i, \ldots, x^N)\right\}^2 \\ &= \frac{1}{N^2} \sup_{x^i, \tilde{x}^i \in \mathcal{X}} \left\{ \left[T(x^i)T(x^i)^T - T(\tilde{x}^i)T(\tilde{x}^i)^T\right] - \frac{1}{N-1}\sum_{j \neq i} \left(T(x^i) - T(\tilde{x}^i)\right) T(x^j)^T ... \right. \\ &\qquad \left. - \frac{1}{N-1}\sum_{j \neq i} T(x^j)\left(T(x^i) - T(\tilde{x}^i)\right)^T \right\}^2 \\ &\leq \frac{1}{N^2} C, \end{aligned}$$

where C is a fixed positive semidefinite matrix. This last inequality is due to Assumption 1(iv) that $T(x)$ is bounded on $\mathcal{X}$. Note that conditioning on $\mathcal{F}_{k-1}$, $\{x_k^i, i = 1, \ldots, N_k\}$ are i.i.d., and $E_{\theta_k}[\widehat{V}_k|\mathcal{F}_{k-1}] = V_k$. Then according to the matrix bounded differences inequality, for all $\delta > 0$,

$$P\left\{\lambda_{\max}(\widehat{V}_k - V_k) \geq \delta \mid \mathcal{F}_{k-1}\right\} \leq d \exp\left(\frac{-N_k\delta^2}{8\|C\|_2}\right),$$

which also implies

$$P\left\{-\lambda_{\min}(\widehat{V}_k - V_k) \geq \delta \mid \mathcal{F}_{k-1}\right\} = P\left\{\lambda_{\max}(V_k - \widehat{V}_k) \geq \delta \mid \mathcal{F}_{k-1}\right\} \leq d \exp\left(\frac{-N_k\delta^2}{8\|C\|_2}\right).$$

Recall that for a symmetric matrix A, $\|A\|_2 = \max(\lambda_{\max}(A), -\lambda_{\min}(A))$ and $\|A\| \leq \|A\|_2$. Hence,

$$P\left\{\|\widehat{V}_k - V_k\| \geq \delta \mid \mathcal{F}_{k-1}\right\} \leq P\left\{\|\widehat{V}_k - V_k\|_2 \geq \delta \mid \mathcal{F}_{k-1}\right\} \leq 2d \exp\left(\frac{-N_k\delta^2}{8\|C\|_2}\right).$$

Recall that for any nonnegative random variable X,

$$\begin{aligned} E[X] &= \int_0^\infty P(X \geq x)dx \\ &\leq a + \int_a^\infty P\left(X \geq x\right) dx. \end{aligned}$$

So we have

$$\begin{aligned} E\left[\|\widehat{V}_k - V_k\|^2 \mid \mathcal{F}_{k-1}\right] &\leq a + \int_a^\infty P\left\{\|\widehat{V}_k - V_k\| \geq \sqrt{x} \mid \mathcal{F}_k\right\} dx \\ &\leq a + \int_a^\infty 2d \exp\left(\frac{-N_k x}{8\|C\|_2}\right) dx. \end{aligned}$$

Set $a = 8\|C\|_2 \log(2d)/N_k$, and we obtain

$$E\left[\|\widehat{V}_k - V_k\| \mid \mathcal{F}_{k-1}\right]^2 \leq E\left[\|\widehat{V}_k - V_k\|^2 \mid \mathcal{F}_{k-1}\right] \leq \frac{8\|C\|_2(1 + \log(2d))}{N_k}. \tag{28}$$

To bound the second term in the right-hand-side of (26), notice that $\widetilde{E}_{p_k}[T_j(X)]$ is a self-normalized importance sampling estimator of $E_{p_k}[T_j(X)]$, where $T_j(X)$ is the j^{th} element in the vector $T(X)$. Applying Theorem 9.1.10 (pp. 294, [2]), we have

$$E\left[|\widetilde{E}_{p_k}[T_j(X)] - E_{p_k}[T_j(X)]|^2 | \mathcal{F}_{k-1}\right] \leq \frac{c_j}{N_k}, \quad j = 1, \ldots, d,$$

where c_j's are positive constants due to the boundedness of $T_j(x)$ on $\mathcal{X}$. Hence,

$$\begin{aligned} & E\left[\|\widetilde{E}_{p_k}[T(X)] - E_{p_k}[T(X)]\||\mathcal{F}_{k-1}\right]^2 \\ \leq\ & E\left[\|\widetilde{E}_{p_k}[T(X)] - E_{p_k}[T(X)]\|^2|\mathcal{F}_{k-1}\right] \\ \leq\ & \sum_{j=1}^{d} E\left[|\widetilde{E}_{p_k}[T_j(X)] - E_{p_k}[T_j(X)]|^2|\mathcal{F}_{k-1}\right] \leq \frac{d\max_j c_j}{N_k}. \end{aligned} \tag{29}$$

Putting (28) and (29) together, we obtain

$$\begin{aligned} E[\|\xi_k\|] &\leq E\left[\epsilon^{-2}\|\widehat{V}_k - V_k\|\|\widetilde{E}_{p_k}[T(X)] - E_{\theta_k}[T(X)]\| + \|V_k^{-1}\|\|\widetilde{E}_{p_k}[T(X)] - E_{p_k}[T(X)]\|\right] \\ &\leq M\epsilon^{-2}E\left[E\left[\|\widehat{V}_k - V_k\||\mathcal{F}_{k-1}\right]\right] + \epsilon^{-1}E\left[E\left[\|\widetilde{E}_{p_k}[T(X)] - E_{p_k}[T(X)]\||\mathcal{F}_{k-1}\right]\right] \\ &\leq \frac{M\epsilon^{-2}\sqrt{8\|C\|_2(1+\log(2d))} + \epsilon^{-1}\sqrt{d\max_j c_j}}{\sqrt{N_k}} \\ &\triangleq \frac{c}{\sqrt{N_k}}, \end{aligned}$$

where the positive constant M is due to the boundedness of $T(x)$ on $\mathcal{X}$.

Therefore, for any $T > 0$

$$\begin{aligned} E\left[\sum_{i=k}^{\infty}\alpha_i\|\xi_i\|\right] &= \sum_{i=k}^{\infty}\alpha_i E[\|\xi_i\|] \\ &\leq c\sum_{i=k}^{\infty}\frac{\alpha_i}{\sqrt{N_i}} \\ &= c\sum_{i=k}^{\infty}\frac{1}{i^\beta} \\ &\leq c\left(\frac{1}{k^\beta} + \int_k^\infty \frac{1}{x^\beta}dx\right) \\ &= c\left(\frac{1}{k^\beta} + \frac{1}{\beta-1}\frac{1}{k^{\beta-1}}\right), \end{aligned}$$

where the first line follows from the monotone convergence theorem, and the third line follows from Assumption 1(ii). For any $\tau > 0$, we have from Markov's inequality

$$\begin{aligned} P\left(\sum_{i=k}^{\infty}\alpha_i\|\xi_i\| \geq \tau\right) &\leq \frac{E\left[\sum_{i=k}^{\infty}\alpha_i\|\xi_i\|\right]}{\tau} \\ &\leq \frac{c}{\tau}\left(\frac{1}{k^\beta} + \frac{1}{\beta-1}\frac{1}{k^{\beta-1}}\right) \ \to\ 0 \ \text{ as } k\to\infty, \end{aligned}$$

where the last statement is due to $\beta > 1$. This result of convergence in probability together with the fact that the sequence $\{\sum_{i=k}^{\infty} \alpha_i \|\xi_i\|\}$ is monotone implies that the sequence $\{\sum_{i=k}^{\infty} \alpha_i \|\xi_i\|\}$ converges w.p.1 as $k \to \infty$. Furthermore, since $\sup_{\{n:0\le\sum_{i=k}^{n-1}\alpha_i\le T\}} \| \sum_{i=k}^{n} \alpha_i \xi_i\| \le \sup_{\{n:0\le\sum_{i=k}^{n-1}\alpha_i\le T\}} \sum_{i=k}^{n} \alpha_i \|\xi_i\| \le \sum_{i=k}^{\infty} \alpha_i \|\xi_i\|$, we conclude that $\{\sup_{\{n:0\le\sum_{i=k}^{n-1}\alpha_i\le T\}} \| \sum_{i=k}^{n} \alpha_i \xi_i\|\}$ converges to 0 w.p.1 as $k \to \infty$. □

Proof. **Proof of Theorem 1.** To show our theorem, we apply Theorem 2.1 in [11]. The condition on the step size sequence in their theorem is satisfied by our Assumption 1(i), and condition (2.2) there is a result of Lemma 2. Thus, to establish convergence, it is sufficient to show $b_k \to 0$ w.p.1 as $k \to \infty$. Note that

$$\begin{aligned} b_k &= \widehat{V}_k^{-1} \left(\widehat{E}_{p_k}[T(X)] - \widetilde{E}_{p_k}[T(X)] \right) \\ &= \widehat{V}_k^{-1} \left(\frac{\bar{\mathbb{U}}_k}{\bar{\mathbb{V}}_k} - \frac{\bar{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k} + \frac{\bar{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k} - \frac{\tilde{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k} \right) \\ &= \widehat{V}_k^{-1} \bar{\mathbb{U}}_k \Big(\frac{\tilde{\mathbb{V}}_k - \bar{\mathbb{V}}_k}{\bar{\mathbb{V}}_k \tilde{\mathbb{V}}_k} \Big) + \widehat{V}_k^{-1} \frac{\bar{\mathbb{U}}_k - \tilde{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k}. \end{aligned}$$

Hence,

$$\begin{aligned} \|b_k\| &\le \frac{\|\widehat{V}_k^{-1}\| \|\bar{\mathbb{U}}_k\|}{|\bar{\mathbb{V}}_k \tilde{\mathbb{V}}_k|} |\tilde{\mathbb{V}}_k - \bar{\mathbb{V}}_k| + \frac{\|\widehat{V}_k^{-1}\|}{|\tilde{\mathbb{V}}_k|} \|\bar{\mathbb{U}}_k - \tilde{\mathbb{U}}_k\| \\ &\le \frac{\|\widehat{V}_k^{-1}\| \|\bar{\mathbb{U}}_k\|}{|\bar{\mathbb{V}}_k \tilde{\mathbb{V}}_k|} \frac{1}{N_k} \sum_{i=1}^{N_k} |\widehat{S}_{\theta_k}(H(x_k^i)) - S_{\theta_k}(H(x_k^i))| \\ &\quad + \frac{\|\widehat{V}_k^{-1}\|}{|\tilde{\mathbb{V}}_k|} \frac{1}{N_k} \sum_{i=1}^{N_k} |\widehat{S}_{\theta_k}(H(x_k^i)) - S_{\theta_k}(H(x_k^i))| \|T(x_k^i)\|. \end{aligned}$$

Since $T(x)$ is bounded, it is easy to see that $\frac{\|\bar{\mathbb{U}}_k\|}{|\bar{\mathbb{V}}_k|}$ is also bounded. Furthermore, note that $\|\widehat{V}_k^{-1}\|$ is bounded and $|\tilde{\mathbb{V}}_k|$ is bounded away from zero. This together with Assumption 1(iv) imply that the sequence $\{b_k\}$ converges to zero w.p.1. □

Proof. **Proof of Lemma 3.** Under Assumption 1, we know that the sequence $\{\theta_k\}$ converges w.p.1. to a limiting point θ^*. This, together with Assumption 1(iii), implies that the sequence of sampling distributions $\{f(x;\theta_k)\}$ will converge point-wise in x to a limiting distribution $f(x;\theta^*)$ w.p.1. Note that $\|\widehat{\mathrm{Var}}_{\theta_k}(T(X)) - \mathrm{Var}_{\theta^*}(T(X))\| \le \|\widehat{\mathrm{Var}}_{\theta_k}(T(X)) - \mathrm{Var}_{\theta_k}(T(X))\| + \|\mathrm{Var}_{\theta_k}(T(X)) - \mathrm{Var}_{\theta^*}(T(X))\|$. Clearly, the first term converges to zero by the strong consistency of the variance estimator. On the other hand, using the point-wise convergence of $\{f(\cdot;\theta_k)\}$ and the dominated convergence theorem, it is easy to see that the second term also vanishes to zero. This shows $\Phi_k \to \Phi$ w.p.1. Thus,

the convergence of Γ_k to Γ is a direct consequence of the continuity assumption of $J_{\mathcal{L}}$ in the neighborhood of θ^*. Regarding T_k, we have

$$\begin{aligned} T_k &= k^{\tau/2}\Phi_k\Big(\frac{\bar{\mathbb{U}}_k}{\bar{\mathbb{V}}_k} - \frac{\bar{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k} + \frac{\bar{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k} - \frac{\tilde{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k}\Big) + k^{\tau/2}\Phi_k\Big(E_{\theta_k}\Big[\frac{\tilde{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k}\Big|\mathcal{F}_{k-1}\Big] - \frac{\mathbb{U}_k}{\mathbb{V}_k}\Big) \\ &= T_{k,1} + T_{k,2}, \end{aligned}$$

where $T_{k,1} = k^{\tau/2}\Phi_k\bar{\mathbb{U}}_k\Big(\frac{\tilde{\mathbb{V}}_k - \bar{\mathbb{V}}_k}{\bar{\mathbb{V}}_k\tilde{\mathbb{V}}_k}\Big) + k^{\tau/2}\Phi_k\frac{\bar{\mathbb{U}}_k - \tilde{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k}$ and $T_{k,2} = k^{\tau/2}\Phi_k\Big(E_{\theta_k}\Big[\frac{\tilde{\mathbb{U}}_k}{\tilde{\mathbb{V}}_k}\Big|\mathcal{F}_{k-1}\Big] - \frac{\mathbb{U}_k}{\mathbb{V}_k}\Big)$. Note that

$$\begin{aligned} \|T_{k,1}\| &\le \|\Phi_k\|\frac{\|\bar{\mathbb{U}}_k\|}{|\bar{\mathbb{V}}_k\tilde{\mathbb{V}}_k|}k^{\tau/2}|\tilde{\mathbb{V}}_k - \bar{\mathbb{V}}_k| + \|\Phi_k\|\frac{1}{|\tilde{\mathbb{V}}_k|}k^{\tau/2}\|\bar{\mathbb{U}}_k - \tilde{\mathbb{U}}_k\| \\ &\le \frac{\|\Phi_k\|\|\bar{\mathbb{U}}_k\|}{|\bar{\mathbb{V}}_k\tilde{\mathbb{V}}_k|}\frac{k^{\tau/2}}{N_k}\sum_{i=1}^{N_k}|\widehat{S}_{\theta_k}(H(x_k^i)) - S_{\theta_k}(H(x_k^i))| \\ &\quad + \frac{\|\Phi_k\|}{|\tilde{\mathbb{V}}_k|}\frac{k^{\tau/2}}{N_k}\sum_{i=1}^{N_k}|\widehat{S}_{\theta_k}(H(x_k^i)) - S_{\theta_k}(H(x_k^i))|\|T(x_k^i)\| \end{aligned} \tag{30}$$

Since $T(x)$ is bounded, it is easy to see that $\frac{\|\bar{\mathbb{U}}_k\|}{|\bar{\mathbb{V}}_k|}$ is also bounded. Furthermore, note that $|\tilde{\mathbb{V}}_k|$ is bounded away from zero. This, together with the boundedness of $\|\Phi_k\|$ and Assumption 3, imply that the right-hand-side of (30) converges to zero w.p.1.

For term $T_{k,2}$, let $\tilde{\mathbb{U}}_k^i$ and $\mathbb{U}_k^i$ be the ith components of $\tilde{\mathbb{U}}_k$ and $\mathbb{U}_k$, respectively. By using a second order two variable Taylor expansion of $\frac{\tilde{\mathbb{U}}_k^i}{\tilde{\mathbb{V}}_k}$ around $\frac{\mathbb{U}_k^i}{\mathbb{V}_k}$, we have

$$\frac{\tilde{\mathbb{U}}_k^i}{\tilde{\mathbb{V}}_k} = \frac{\mathbb{U}_k^i}{\mathbb{V}_k} + \frac{1}{\mathbb{V}_k}(\tilde{\mathbb{U}}_k^i - \mathbb{U}_k^i) - \frac{\mathbb{U}_k^i}{\mathbb{V}_k^2}(\tilde{\mathbb{V}}_k - \mathbb{V}_k) + \frac{\hat{\mathbb{U}}_k^i}{\hat{\mathbb{V}}_k^3}(\tilde{\mathbb{V}}_k - \mathbb{V}_k)^2 - \frac{1}{\hat{\mathbb{V}}_k^2}(\tilde{\mathbb{U}}_k^i - \mathbb{U}_k^i)(\tilde{\mathbb{V}}_k - \mathbb{V}_k),$$

where $\hat{\mathbb{U}}_k^i$ and $\hat{\mathbb{V}}_k$ are on the line segments from $\tilde{\mathbb{U}}_k^i$ to $\mathbb{U}_k^i$ and from $\tilde{\mathbb{V}}_k$ to $\mathbb{V}_k$. Taking conditional expectations at both sides of the above equation, we have

$$\begin{aligned} \Big|E_{\theta_k}\Big[\frac{\tilde{\mathbb{U}}_k^i}{\tilde{\mathbb{V}}_k}\Big|\mathcal{F}_{k-1}\Big] - \frac{\mathbb{U}_k^i}{\mathbb{V}_k}\Big| &\le E_{\theta_k}\Big[\frac{|\hat{\mathbb{U}}_k^i|}{|\hat{\mathbb{V}}_k^3|}(\tilde{\mathbb{V}}_k - \mathbb{V}_k)^2\Big|\mathcal{F}_{k-1}\Big] + E_{\theta_k}\Big[\frac{1}{\hat{\mathbb{V}}_k^2}|(\tilde{\mathbb{U}}_k^i - \mathbb{U}_k^i)(\tilde{\mathbb{V}}_k - \mathbb{V}_k)|\Big|\mathcal{F}_{k-1}\Big] \\ &\le \mathcal{C}_1 E_{\theta_k}\Big[(\tilde{\mathbb{V}}_k - \mathbb{V}_k)^2\Big|\mathcal{F}_{k-1}\Big] + \mathcal{C}_2 E_{\theta_k}\Big[|(\tilde{\mathbb{U}}_k^i - \mathbb{U}_k^i)(\tilde{\mathbb{V}}_k - \mathbb{V}_k)|\Big|\mathcal{F}_{k-1}\Big] \end{aligned} \tag{31}$$

for constants $\mathcal{C}_1 > 0$ and $\mathcal{C}_2 > 0$. Thus, a straightforward calculation shows that the right-hand-side of (31) is $O(N_k^{-1})$. Consequently, we have $T_{k,2} \to 0$ w.p.1. as $k \to \infty$ by taking $N_k = N_0 k^{\zeta}$ with $\zeta > \tau/2$. This shows $T_k \to 0$ w.p.1. as desired. □

Proof. **Proof of Lemma 4.** $E_{\theta_k}[W_k|\mathcal{F}_{k-1}] = 0$ follows directly from the definition of W_k. Again, we let $\tilde{\mathbb{U}}_k^i$ and $\mathbb{U}_k^i$ be the ith components of $\tilde{\mathbb{U}}_k$ and $\mathbb{U}_k$, let $T_i(x)$ be the ith component of the sufficient statistic $T(x)$, and define $\Sigma_{i,j}^k$ as the (i,j)th entry of the matrix $E_{\theta_k}[W_k W_k^T|\mathcal{F}_{k-1}]$. By using a first order two variable Taylor expansion of $\frac{\tilde{\mathbb{U}}_k^i}{\tilde{\mathbb{V}}_k}$ around $\frac{\mathbb{U}_k^i}{\mathbb{V}_k}$, we have

$$\frac{\tilde{\mathbb{U}}_k^i}{\tilde{\mathbb{V}}_k} = \frac{\mathbb{U}_k^i}{\mathbb{V}_k} + \frac{1}{\mathbb{V}_k}(\tilde{\mathbb{U}}_k^i - \mathbb{U}_k^i) - \frac{\mathbb{U}_k^i}{\mathbb{V}_k^2}(\tilde{\mathbb{V}}_k - \mathbb{V}_k) + \mathcal{R}_k, \tag{32}$$

where $\mathcal{R}_k$ is a reminder term. Therefore, $\Sigma_{i,j}^k$ can be expressed as

$$\begin{aligned}
\Sigma_{i,j}^k =& k^{\tau-\alpha} E_{\theta_k}\Big[\Big(\frac{\tilde{\mathbb{U}}_k^i}{\tilde{\mathbb{V}}_k} - E_{\theta_k}\Big[\frac{\tilde{\mathbb{U}}_k^i}{\tilde{\mathbb{V}}_k}\Big|\mathcal{F}_{k-1}\Big]\Big)\Big(\frac{\tilde{\mathbb{U}}_k^j}{\tilde{\mathbb{V}}_k} - E_{\theta_k}\Big[\frac{\tilde{\mathbb{U}}_k^j}{\tilde{\mathbb{V}}_k}\Big|\mathcal{F}_{k-1}\Big]\Big)\Big|\mathcal{F}_{k-1}\Big] \\
=& k^{\tau-\alpha}\frac{1}{\mathbb{V}_k^2} E_{\theta_k}[(\tilde{\mathbb{U}}_k^i - \mathbb{U}_k^i)(\tilde{\mathbb{U}}_k^j - \mathbb{U}_k^j)|\mathcal{F}_{k-1}] \quad [i] \\
&- k^{\tau-\alpha}\frac{\mathbb{U}_k^j}{\mathbb{V}_k^3} E_{\theta_k}[(\tilde{\mathbb{U}}_k^i - \mathbb{U}_k^i)(\tilde{\mathbb{V}}_k - \mathbb{V}_k)|\mathcal{F}_{k-1}] \quad [ii] \\
&- k^{\tau-\alpha}\frac{\mathbb{U}_k^i}{\mathbb{V}_k^3} E_{\theta_k}[(\tilde{\mathbb{U}}_k^j - \mathbb{U}_k^j)(\tilde{\mathbb{V}}_k - \mathbb{V}_k)|\mathcal{F}_{k-1}] \quad [iii] \\
&+ k^{\tau-\alpha}\frac{\mathbb{U}_k^i \mathbb{U}_k^j}{\mathbb{V}_k^4} E_{\theta_k}[(\tilde{\mathbb{V}}_k - \mathbb{V}_k)^2|\mathcal{F}_{k-1}] \quad [iv] \\
&+ k^{\tau-\alpha}\bar{\mathcal{R}}_k,
\end{aligned}$$

where $\bar{\mathcal{R}}_k$ represents a higher-order term.

$$\begin{aligned}
[i] &= k^{\tau-\alpha}\frac{1}{\mathbb{V}_k^2}\Big(E_{\theta_k}[\tilde{\mathbb{U}}_k^i\tilde{\mathbb{U}}_k^j|\mathcal{F}_{k-1}] - \mathbb{U}_k^i\mathbb{U}_k^j\Big) \\
&= k^{\tau-\alpha}\frac{1}{\mathbb{V}_k^2}\frac{1}{N_k}\Big(E_{\theta_k}\big[S_{\theta_k}^2(H(X))T_i(X)T_j(X)\big|\mathcal{F}_{k-1}\big] - \mathbb{U}_k^i\mathbb{U}_k^j\Big) \\
&= k^{\tau-\alpha}\frac{1}{N_k}\Big(\frac{E_{\theta_k}\big[S_{\theta_k}^2(H(X))T_i(X)T_j(X)\big|\mathcal{F}_{k-1}\big]}{E_{\theta_k}^2[S_{\theta_k}(H(X))]} - \frac{\mathbb{U}_k^i\mathbb{U}_k^j}{\mathbb{V}_k^2}\Big) \\
&= \frac{k^{\tau-\alpha}}{N_k}\Big[E_{p_k}\Big[T_i(X)T_j(X)\frac{p_k(X)}{f(X;\theta_k)}\Big] - E_{p_k}[T_i(X)]E_{p_k}[T_j(X)]\Big].
\end{aligned}$$

By using a similar argument, it can be seen that

$$\begin{aligned}
[ii] &= \frac{k^{\tau-\alpha}}{N_k}\Big[E_{p_k}\big[T_j(X)\big]E_{p_k}\Big[T_i(X)\frac{p_k(X)}{f(X;\theta_k)}\Big] - E_{p_k}[T_i(X)]E_{p_k}[T_j(X)]\Big], \\
[iii] &= \frac{k^{\tau-\alpha}}{N_k}\Big[E_{p_k}\big[T_i(X)\big]E_{p_k}\Big[T_j(X)\frac{p_k(X)}{f(X;\theta_k)}\Big] - E_{p_k}[T_i(X)]E_{p_k}[T_j(X)]\Big], \\
[iv] &= \frac{k^{\tau-\alpha}}{N_k}\Big[E_{p_k}\big[T_j(X)\big]E_{p_k}\big[T_i(X)\big]E_{p_k}\Big[\frac{p_k(X)}{f(X;\theta_k)}\Big] - E_{p_k}[T_i(X)]E_{p_k}[T_j(X)]\Big].
\end{aligned}$$

Therefore,

$$\begin{aligned}\Sigma_{i,j}^k &= [i] - [ii] - [iii] + [iv] + k^{\tau-\alpha}\bar{\mathcal{R}}_k \\ &= \frac{k^{\tau-\alpha}}{N_k} E_{p_k}\Big[(T_i(X) - E_{p_k}[T_i(X)])(T_j(X) - E_{p_k}[T_j(X)])\frac{p_k(X)}{f(X;\theta_k)}\Big] + k^{\tau-\alpha}\bar{\mathcal{R}}_k \\ &= \frac{k^{\tau-\alpha}}{N_k} E_{\theta_k}\Big[(T_i(X) - E_{p_k}[T_i(X)])(T_j(X) - E_{p_k}[T_j(X)])\frac{p_k^2(X)}{f^2(X;\theta_k)}\Big] + k^{\tau-\alpha}\bar{\mathcal{R}}_k.\end{aligned}$$

By taking $N_k = N_0 k^{\tau-\alpha}$, it can be shown that the higher-order term $k^{\tau-\alpha}\bar{\mathcal{R}}_k$ is $o(1)$. In addition, since $S_\theta(y)$ is continuous in θ for a fixed y, the point-wise convergence of $f(\cdot;\theta_k)$ to $f(\cdot;\theta^*)$ implies that $p_k(x)$ will also converge in a point-wise manner to a limiting distribution $p_*(x)$. Thus, the dominated convergence theorem suggests that $\Sigma_{i,j}^k$ will converge to

$$\Sigma_{i,j} = \mathcal{C} E_{\theta^*}\Big[(T_i(X) - E_{p_*}[T_i(X)])(T_j(X) - E_{p_*}[T_j(X)])\frac{p_*^2(X)}{f^2(X;\theta^*)}\Big]$$

for some positive constant $\mathcal{C}$. Therefore, the limiting matrix Σ is given by

$$\Sigma = \mathrm{Cov}_{\theta^*}\Big((T(X) - E_{p_*}[T(X)])\frac{p_*(X)}{f(X;\theta^*)}\Big),$$

where $\mathrm{Cov}_{\theta^*}(\cdot)$ is the covariance matrix with respect to $f(\cdot;\theta^*)$.

To show the last statement, we use Hölder's inequality and write

$$\lim_{k\to\infty} E[I\{\|W_k\|^2 \geq rk^\alpha\}\|W_k\|^2] \leq \limsup_{k\to\infty}\Big[P\big(\|W_k\|^2 \geq rk^\alpha\big)\Big]^{\frac{1}{2}}\Big[E\big[\|W_k\|^4\big]\Big]^{\frac{1}{2}}. \tag{33}$$

Note that

$$\begin{aligned}P\big(\|W_k\|^2 \geq rk^\alpha\big) &= P\big(\|W_k\| \geq \sqrt{r}k^{\alpha/2}\big) \\ &\leq \frac{E[\|W_k\|^2]}{rk^\alpha} \quad \text{by Chebyshev's inequality} \\ &= \frac{E\big[E_{\theta_k}[\|W_k\|^2|\mathcal{F}_{k-1}]\big]}{rk^\alpha} \\ &= \frac{E\big[\mathrm{tr}(\Sigma^k)\big]}{rk^\alpha} \\ &= O(k^{-\alpha})\end{aligned}$$

by taking $N_k = N_0 k^{\tau-\alpha}$ for k sufficiently large, where the last step follows because all entries in Σ^k are bounded and thus convergence w.p.1. implies convergence in expectation. On the other hand, by (32), $E[\|W_k\|^4]$ can be expressed in terms of the fourth order central moments of the sample mean and it can be verified that $E[\|W_k\|^4] = O(1)$. This shows that the right-hand-side of (33) is $O(k^{-\frac{\alpha}{2}})$, which vanishes to zero as $k \to \infty$. □

Acknowledgments: The authors gratefully acknowledge the support by the National Science Foundation under Grants ECCS-0901543 and CMMI-1130273 and Air Force Office of Scientific Research under YIP Grant FA-9550-12-1-0250. We are grateful to Xi Chen, graduate student in the Department of Industrial & Enterprise Systems Engineering at UIUC, for her help with conducting the numerical experiments in Section 5.

References

[1] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[2] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, 2005.

[3] M. Dorigo and C. Blum. Ant colony optimization theory: a survey. *Theoretical Computer Science*, 344:243 – 278, 2005.

[4] M. Dorigo and L.M. Gambardella. Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, 1:53 – 66, 1997.

[5] V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327 – 1332, 1968.

[6] F. W. Glover. Tabu search: A tutorial. *Interfaces*, 20:74 – 94, 1990.

[7] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1989.

[8] J. Hu, M. C. Fu, and S. I. Marcus. A model reference adaptive search method for global optimization. *Operations Research*, 55:549–568, 2007.

[9] J. Hu, P. Hu, and H. S. Chang. A stochastic approximation framework for a class of randomized optimization algorithms. *IEEE Transactions on Automatic Control*, 57(1):165–178, 2012.

[10] S. Kirkpatrick, C. D. Gelatt, and Jr. M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[11] H. J. Kushner. Stochastic approximation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):87–96, 2010.

[12] H. J. Kushner and D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York, NY, 1978.

[13] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York, NY, 2nd edition, 2004.

[14] P. Larranaga, R. Etxeberria, J. A. Lozano, and J. M. Pena. Optimization by learning and simulation of Bayesian and Gaussian networks. Technical Report EHU-KZAA-IK-4/99, Department of Computer Science and Artificial Intelligence, University of the Basque Country, 1999.

[15] C. McDiarmid. *Surveys in Combinatorics*, chapter On the Method of Bounded Differences, pages 148 – 188. Cambridge University Press, Cambridge, 1989.

[16] Sean Meyn. Variance in stochastic approximation. Note for private communication, 2009.

[17] O. Molvalioglu, Z. B. Zabinsky, and W. Kohn. The interacting-particle algorithm with dynamic heating and cooling. *Journal of Global Optimization*, 43:329–356, 2009.

[18] O. Molvalioglu, Z. B. Zabinsky, and W. Kohn. Meta-control of an interacting-paricle algorithm. *Nonlinear Analysis: Hybrid Systems*, 4(4):659 – 671, 2010.

[19] H. Muhlenbein and G. Paaß. From recombination of genes to the estimation of distributions: I. binary parameters. In H. M. Voigt, W. Ebeling, I. Rechenberg, and H. P. Schwefel, editors, *Parallel Problem Solving from Nature-PPSN IV*, pages 178–187, Berlin, Germany, 1996. Springer Verlag.

[20] B. Polyak. New stochastic approximation type procedures. *Automation and Remote Control*, 51:937–946, 1990.

[21] C.R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

[22] H. E. Romeijn and R. L. Smith. Simulated annealing for constrained global optimization. *Journal of Global Optimization*, 5(2):101–126, 1994.

[23] R. Y. Rubinstein. Combinatorial optimization, ants and rare events. In S. Uryasev and P.M. Pardalos, editors, *Stochastic Optimization: Algorithms and Applications*, pages 304–358, Dordrecht, The Netherlands, 2001. Kluwer Academic Publishers.

[24] D. Ruppert. Stochastic approximation. In B.K. Ghosh and P.K. Sen, editors, *Handbook in Sequential Analysis*, page 503 529. Marcel Dekker, New York, 1991.

[25] L. Shi and S. Ólafsson. Nested partitions method for global optimization. *Operations Research*, 48(3):390 – 407, 2000.

[26] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics, Aug. 2011, .*, 2011. doi:10.1007/s10208-011-9099-z.

[27] D. H. Wolpert. Finding bounded rational equilibria part i: Iterative focusing. In T. Vincent, editor, *Proceedings of the Eleventh International Symposium on Dynamic Games and Applications*, 2004.

[28] Z. B. Zabinsky. *Stochastic Adaptive Search for Global Optimization*. Nonconvex Optimization and Its Applications. Springer, 2003.

[29] E. Zhou and X. Chen. Sequential Monte Carlo simulated annealing. *Journal of Global Optimization*, 2011. Under review.

[30] M. Zlochin, M. Birattari, N. Meuleau, and M. Dorigo. Model-based search for combinatorial optimization: A critical survey. *Annals of Operations Research*, 131:373–395, 2004.

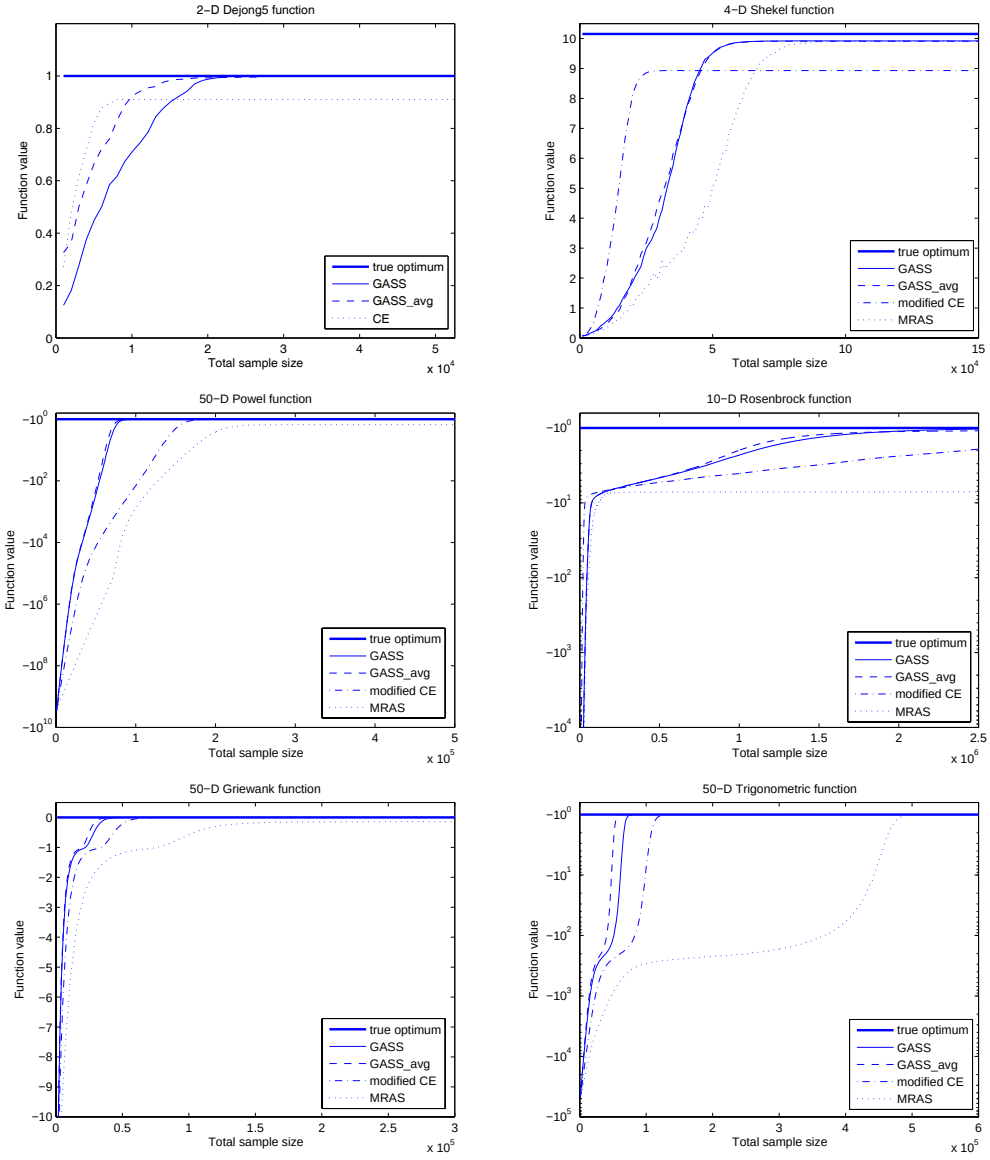

Figure 1: Comparison of GASS, GASS_avg, modified CE and MRAS

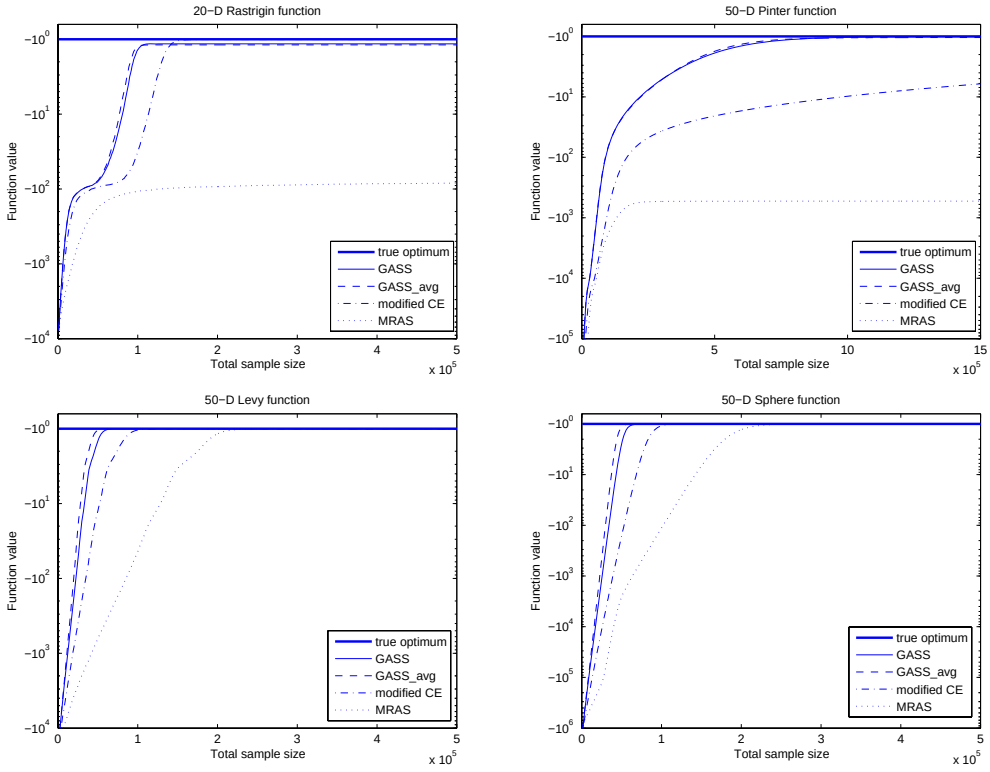

Figure 2: Comparison of GASS, GASS_avg, modified CE and MRAS